Julia Korngiebel

Vergleichsarbeiten und ihr Potenzial für die Schul- und Unterrichtsentwicklung.

Eine qualitative Untersuchung zur Nutzung der Lernstandserhebungen an hessischen Gymnasien.

Dissertation

Philipps-Universität Marburg
Fachbereich Erziehungswissenschaften
Hochschulkennziffer: 1180

Geburtsort: Eisenach
Gutachter: Prof. Dr. Rainer Lersch
Prof. Dr. Heike Ackermann

Einreichungstermin: 08.10.2013
Prüfungstermin: 27.03.2014

Marburg/Lahn, 2014

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CCC	Category Charakteristic Curve
DESI	Deutsch-Englisch-Schülerleistungen-International
EMSE	Netzwerk Empiriegestützte Schulentwicklung
GER	Gemeinsamer Europäischer Referenzrahmen für Sprachen
ICC	Item Characteristic Curve
IGLU	Internationale Grundschul-Lese-Untersuchung
IQB	Institut zur Qualitätsentwicklung im Bildungswesen
KMK	Ständige Konferenz der Kultusminister der Länder (Kultusministerkonferenz)
LAU	Aspekte der Lernausgangslage und der Lernentwicklung
LSA	Landesschulamt und Lehrkräfteakademie, Wiesbaden
MARKUS	Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
QuaSUM	Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik
SINUS	Steigerung der Effizienz des mathematischen und naturwissenschaftlichen Unterrichts (Modellprojekt)
TIMSS	Third International Mathematics and Science Study
VERA 3	Vergleichsarbeiten in der Klassenstufe 3
VERA 6	Vergleichsarbeiten in der Klassenstufe 6
VERA 8	Vergleichsarbeiten in der Klassenstufe 8
VERA	VERgleichsArbeiten in der Grundschule (ehemaliges Testkonzept in Klassenstufe 4)

1 Einleitung

1.1 Einführung in die Thematik

Seit der Veröffentlichung der Ergebnisse zur PISA-Untersuchung, welche dem Bildungssystem der Bundesrepublik Deutschleid defizitäre *Schülerleistungen* in einigen Lernfeldern attestierten, ist bereits mehr als ein Jahrzehnt vergangen. Die öffentliche Debatte über die Resultate dieser Leistungsmessung ist mittlerweile abgeebbt. Die bildungspolitischen Nachwirkungen in Form von vielfältigen Reformansätzen im Bildungswesen haben jedoch den schulischen Alltag nachhaltig verändert.

Es entstand unter anderem die Forderung, die Fähig- und Fertigkeiten der Schüler[1] mithilfe von standardisierten Tests zunehmend zu verorten und vergleichend bewerten zu können. Mehrere Beschlüsse zur Teilnahme an verschiedenen Leistungstests auf Bundes-, Landes- beziehungsweise Schulebene bekräftigen dies.

Des Weiteren stellte sich in einer wissenschaftlichen Diskussion die Fragestellung heraus, wie Schülerleistungen in standardisierten Tests präzise gemessen werden können und welche Fähigkeiten der Lernenden dabei als erstrebenswert zu erachtet sind. Eine Reaktion zu dieser Erörterung war die Implementierung von Bildungsstandards, in denen die geforderten Fähig- und Fertigkeiten präzisiert werden. Die Bildungsstandards umfassen eine konkrete Formulierung von *Kompetenzen*, über welche die Schüler bis zum Ende der wesentlichen Abschnitte ihrer Schullaufbahn tatsächlich verfügen sollen.

Diese beiden in sehr groben Ansätzen skizzierten Entwicklungen - die Forderung nach zunehmender Ergebnismessung sowie die Einführung von Bildungsstandards - wurden miteinander verknüpft in einem weiteren Baustein des bildungspolitischen Maßnahmenpakets. Der Erwerb spezifischer Kompetenzen entsprechend der Bildungsstandards erfordert eine Überprüfung der Ausprägung der intendierten Kompetenz zu regelmäßigen Zeitpunkten in der Schullaufbahn eines Schülers. Auf dieser Grundlage kann in einem kontinuierlichen Prozess der Weg des Kompetenzerwerbs zielgerichtet gestaltet werden. Neben vielfältigen Diagnosemöglichkeiten des Unterrichts erfolgt dies zusätzlich durch extern entwickelte Vergleichsarbeiten, die von den Lehrkräften sowohl die Orientierung an den Bildungsstandards bei der Unterrichtsgestaltung als auch die Überprüfung der erreichten Kompetenzen einfordern.

[1] Da die Verwendung der jeweils beiderseitigen Geschlechterbezeichnung als unverhältnismäßig umständlich erscheint, soll es in der nachfolgenden Abhandlung ausreichen, das männliche Genus zu benennen. Selbstverständlich sind aber stets die Vertreterinnen und Vertreter beider Geschlechter gemeint.

Die Vergleichsarbeiten sind demnach externe, standardisierte Leistungsmessungen, welche die Kompetenzen der Schüler zu in der Primar- und Sekundarstufe I messen und einen Vergleich mit den durchschnittlichen Resultaten des jeweiligen Bundeslandes ermöglichen. Die Überprüfung der Fähig- und Fertigkeiten in verschiedenen Kompetenzbereichen wird wissenschaftlich ausgewertet. Das Testkonzept ist primär an die schulischen Akteure adressiert. Indem die Lehrkraft Aufschluss über den momentanen Leistungsstand erhält, kann sie die Stärken der Schüler durch geeignete Maßnahmen fördern und die Schwächen reduzieren. Auf diese Weise sollen die Vergleichsarbeiten die *Schul- und Unterrichtsentwicklung* forcieren, so dass sie langfristig einen Beitrag zu einer Verbesserung der Schülerleistungen erbringen.

Ein aufwändig konzipiertes Testinstrument ist allerdings wirkungslos, wenn seine intendierten Zielsetzungen in der Praxis nicht erreicht werden. Eine Untersuchung der Differenz zwischen den Funktionen und den tatsächlichen Effekten der Vergleichsarbeiten erfordert die Beobachtung zu der Verwendung der Tests auf der Ebene der individuellen Schule, in der die zentralen Prozesse wie die Testdurchführung, die Bewertung der Testergebnisse sowie die in Reaktion auf die Resultate ergriffenen Handlungen durch die Lehrkräfte vorgenommen werden. Die Herausforderung des Umgangs mit dieser neuartigen Form der Leistungsmessung stellt sich aus diesem Grund vorrangig den Lehrpersonen sowie übergeordnet den Schulleitungen. Mit jeder erneuten Durchführung des Tests schreitet der Erfahrungsprozess voran, so dass die beteiligten Lehrkräfte zu beurteilenden Einschätzungen bezüglich der folgenden Aspekte gelangen:

- Konzeption der Vergleichsarbeiten (Inhalt, Umfang und Gestaltung),
- Prozess der Durchführung der Vergleichsarbeiten,
- Mehrwert der Vergleichsarbeiten für die eigene schulische Arbeit sowie
- individuelle Bewertung des Nutzen-Aufwand-Verhältnisses.

Die Bewertungen dieser Gesichtspunkte durch die Lehrkräfte und Schulleitungen sind in einigen Bundesländern bislang nur in Ansätzen und in anderen noch nicht untersucht worden.

Zu den Einschätzungen der schulischen Akteure an den Nutzen der Vergleichsarbeiten treten die übergeordneten Erwartungshaltungen der bildungspolitischen Institutionen hinzu, welche die Verantwortung über die Vergleichsarbeiten in ihrem jeweiligen Bundesland innehaben. Deren Perspektive richtet sich an der Zielsetzung aus, die Nutzung der Tests in den Schulen in besonderem Maße zu befördern, so dass die anvisierten Effekte tatsächlich zu einer Qualitätssicherung und -entwicklung der Schulentwicklung beitragen.

Vor diesem Hintergrund sollte zum einen die Fragestellung thematisiert werden, was die Vergleichsarbeiten überhaupt bewirken *können*. Zum anderen ist die Realisierung dieser festgelegten Ziele anhand der Testverwendung im schulischen Alltag zu überprüfen. Eine umfassende Analyse dieses Wirkungskomplexes erfordert zugleich die Erforschung der relevanten Einflussfaktoren auf den Nutzungsprozess.

In diesem Zusammenhang wurde innerhalb der Schul-und Unterrichtsforschung als Teilgebiet der Bildungsforschung eine qualitative Interviewstudie mit Lehrkräften und Mitgliedern der Schulleitung an zwölf hessischen Gymnasien zur Nutzung der Vergleichsarbeiten (in Hessen als Lernstandserhebungen bezeichnet) durchgeführt. Die Ergebnisse aus dieser Untersuchung wurden mithilfe der qualitativen Inhaltsanalyse ausgewertet, welche eine Beantwortung der folgenden grundlegenden Fragestellungen dieser Arbeit ermöglichte.

1.2 Fragestellungen der Arbeit

Fragestellung 1: Welches Potenzial bieten die Vergleichsarbeiten für die Schulentwicklung?

Wie bereits erläutert wurde, liegen den Vergleichsarbeiten verschiedene Zielsetzungen zugrunde. Eine Diskussion ihres Potenzials erfordert zunächst eine theoretische Analyse der Funktionen. Verbunden ist hiermit zugleich eine Überprüfung, inwiefern diese durch das entwickelte Testkonzept realisiert werden können oder ob bereits die Gestaltung der Tests, die Regularien zur Durchführung und Auswertung oder die Art der Ergebnisrückmeldung ein Erreichen der intendierten Effekte bereits teilweise ausschließen. Aus dieser Betrachtung heraus kann das tatsächliche Potenzial der Vergleichsarbeiten erschlossen werden.

Fragestellung 2: Inwiefern wird das Potenzial der Vergleichsarbeiten in hessischen Gymnasien für die Schulentwicklung genutzt?

Wenn ein Test das Potenzial besitzt, die Qualitätssicherung und -entwicklung einer Schule im Sinne einer langfristigen Verbesserung der Schülerleistungen zu befördern, bedeutet dies im Umkehrschluss nicht automatisch, dass dieses Potenzial auch genutzt wird. Den Lehrkräften obliegt primär die Verantwortung für die Verwendung der Resultate aus den Vergleichsarbeiten. Dabei verläuft der Nutzungsprozess höchst individuell und unterscheidet sich hinsichtlich der Art der Nutzung, der Nutzungsgegenstände, der -intensität und der -resultate zwischen den einzelnen Lehrkräften erheblich voneinander. Dennoch muss sich ein Testinstrument vor allem an seiner Wirkung messen lassen. Es ist daher bedeutsam zu untersuchen, inwiefern das vorhandene Potenzial der Vergleichsarbeiten bei der Verwendung der Testresultate durch die Lehrkräfte tatsächlich genutzt wurde.

Fragestellung 3: Welche Faktoren greifen fördernd beziehungsweise hemmend in den Nutzungsprozess durch die Lehrkräfte ein?

Die Verwendung der Vergleichsarbeiten kann zum einen zu verschiedenen Zeitpunkten erfolgen, beispielsweise während der Durchführung, der Korrektur oder mit/ nach Erhalt der Testergebnisse. Ebenso stellt die Nutzung selbst einen Prozess dar, der sich in verschiedene aufeinanderfolgende Phasen unterteilen lässt.

Eine intensive und sinnvolle Verwendung der Vergleichsarbeiten gelingt jedoch nicht automatisch mit der Teilnahme an dem Test. Vielmehr beeinflussen zahlreiche Aspekte diesen Prozess. Hierzu zählen beispielsweise die Motivation und die Arbeitsbereitschaft der jeweiligen Lehrperson, die zur Verfügung stehende Arbeitszeit sowie die Testkonzeption selbst. Die vielfältigen Einflussfaktoren haben Auswirkungen auf die Nutzung, indem sie sie im positiven Sinne befördern beziehungsweise hemmen können. Für eine umfassende Beantwortung der Fragestellung 2 ist demnach eine Untersuchung erforderlich, welche Aspekte den Umgang mit den Vergleichsarbeiten in besonderem Maße beeinflussen und welche Auswirkungen auf den Nutzen der Leistungsmessung für die Schulentwicklung dabei festgestellt werden können.

1.3 Aufbau der Arbeit

Die Arbeit gliedert sich in vier Teile:

1. Theoretische Analyse der Vergleichsarbeiten unter Berücksichtigung des aktuellen Forschungsstandes,
2. Erörterung der methodischen Vorgehensweise der Untersuchung,
3. Präsentation der erzielten Ergebnisse aus der Interviewstudie,
4. Beantwortung der Fragestellungen dieser Arbeit.

Der Kontext zur Einführung der Vergleichsarbeiten wurde bislang in Abschnitt 1.1 nur angerissen. Dieser ist jedoch für das umfassende Verständnis der Zielsetzungen der Tests bedeutsam. Daher erfolgt zu Beginn der theoretischen Betrachtungen eine Darlegung zu der Einführung der Bildungsstandards, ihren zugehörigen Kompetenzen und den daraus resultierenden Kerncurricula (vgl. Abschnitt 2). Daran anknüpfend werden in Abschnitt 3 Testmöglichkeiten beschrieben, welche eine standardisierte Messung der vorhandenen Kompetenzen bei den Schülern und somit eine Überprüfung der Bildungsstandards verfolgen.

Zu diesen Formen der Leistungsmessung zählen unter anderem die Vergleichsarbeiten. In Abschnitt 4 werden zunächst die Einführung der Vergleichsarbeiten sowie deren zugrundeliegenden Ziele und Funktionen erläutert. Für die Identifizierung des Potenzials dieser Leistungsmessung in Bezug auf die Schulentwicklung ist es zugleich relevant zu wissen, wie

die Tests konzipiert und durchgeführt werden. Daher wird in diesem Abschnitt des Weiteren der gesamte Prozess ausgehend von der Aufgabenentwicklung, über die Durchführung und der Korrektur bis hin zur Ergebnisrückmeldung und der Evaluation der Vergleichsarbeiten dargelegt. Anschließend werden die Risiken und Grenzen diskutiert, welche mit den Tests verbunden sind (vgl. Abschnitt 4.4).

Im weiteren Verlauf erfolgt eine Betrachtung der Vergleichsarbeiten im Kontext der Schulentwicklung (vgl. Abschnitt 5). Hierzu zählt zum einen die Erläuterung der verschiedenen Bereiche der Schulentwicklung. Zum anderen werden die theoretischen Verwendungsmöglichkeiten der Vergleichsarbeiten innerhalb der Schulentwicklung erörtert. Es findet in diesem Zusammenhang eine vergleichende Erläuterung verschiedener Nutzungs- und Wirkungsmodelle bei der Verwendung der Leistungsmessung sowie eine Typisierung der möglichen Nutzungsformen statt. Die theoretischen Betrachtungen schließen mit einer Darlegung der bisherigen Forschungsergebnisse zu der Verwendung externer Schulleistungsmessungen durch die Lehrkräfte und Schulleitungen, welche auch Untersuchungen zu den Vergleichsarbeiten umfasst (vgl. Abschnitt 5.7).

Im zweiten übergeordneten Teil, den „methodischen Betrachtungen“, werden sowohl das Studiendesign als auch die gewählte Methode zur Auswertung der Untersuchungsergebnisse begründend erläutert (vgl. Abschnitt 6 und 7).

Zu Beginn des Teils „Auswertung der Untersuchung“ erfolgt in Abschnitt 8 eine Charakterisierung der Besonderheiten der Vergleichsarbeiten im Bundesland Hessen, welche dort als Lernstandserhebungen bezeichnet werden. Diese Merkmale betreffen insbesondere die Implementierung der Tests in Hessen, die wissenschaftlich durchgeführte Auswertung der Testergebnisse inklusive des Rückmeldekonzepts für die Lehrkräfte sowie die Kommunikation zwischen den schulischen Akteuren und dem für die Lernstandserhebungen zuständigen Institut.

Anschließend werden in Abschnitt 9 die Ergebnisse der Interviewstudie detailliert präsentiert. Strukturiert wird diese Darlegung nach dem Ablauf eines idealtypischen Nutzungsprozesses, welcher verschiedene Phasen beinhaltet und bereits in Abschnitt 5.5 ausführlich erläutert wird. Die beobachteten Einflussfaktoren auf die Verwendung der Lernstandserhebung von hessischen Gymnasiallehrkräften erfahren ebenfalls eine Erörterung.

Im letzten Teil der Arbeit erfolgt eine knappe Zusammenfassung der wesentlichen Ergebnisse der Untersuchung (vgl. Abschnitt 13). Anschließend werden in Abschnitt 14 die in Abschnitt 1.2 formulierten Fragestellungen der Arbeit beantworte. Die Abhandlung endet anschließend mit einem Ausblick auf den weiteren Forschungsbedarf zu dieser Thematik (vgl. Abschnitt 15).

TEIL A - THEORETISCHE BETRACHTUNGEN

2 Bildungsstandards, Kompetenzmodelle und Kerncurricula

2.1 Von der Input- zur Outputsteuerung

Eine Diskussion über die Qualität von Schule setzte bereits mit den Reformbemühungen der 1970er Jahre ein. Mit der Titulierung einer „Bildungskatastrophe“ entzündete sich die Debatte über das bisherige Verständnis der Qualifikations-, Integrations- und Selektionsfunktion von Schule (vgl. Steffens, 2007, S. 21). Zudem fand eine Auseinandersetzung bezüglich der Neuausrichtung der Systemsteuerung des Bildungswesens statt. Das Reformkonzept, welches 1973 vom Deutschen Bildungsrat vorgelegt wurde, forderte im Zuge der Kritik an der stark bürokratisch-hierarchisch verwalteten Schule eine verstärkte organisatorische Selbstständigkeit der Einzelschule. Dies beinhaltete eine Beschränkung der staatlichen Verwaltung auf Rahmenvorgaben und deren Überprüfung (vgl. Deutscher Bildungsrat, 1973, S. A11). Eine Umsetzung dieser Reformvorstöße gelang jedoch nicht.

Ende der 1980er Jahren wurde die Debatte zur Schulqualität und die Frage nach den Steuerungsmechanismen erneut aufgegriffen. Die Schulen wurden nach Fend (1986) als höchst individuelle Organisationen mit jeweils verschiedener Qualität charakterisiert, so dass die Einzelschule trotz vorgegebener Strukturen individuell mit eigenem Profil als eine pädagogische Handlungseinheit zu agieren habe. Im Zentrum der Auseinandersetzung standen dementsprechend die innere Organisation von Schule sowie der Umgang mit den externen Rahmenbedingungen des Bildungswesens.

Im Zusammenhang mit dem Beginn der Schulqualitätsdiskussion ist zu bemerken, dass die Bundesrepublik bis Ende der 1980er Jahre an keiner internationalen Schulleistungsstudie teilnahm. Das Ziel solcher Messungen ist die Erhebung von Schülerleistungen bestimmter Altersgruppen in international vergleichender Perspektive. Mit den Ergebnissen der Studien können Korrelationen zu charakteristischen Merkmalen der Bildungssysteme und somit Entwicklungspotenziale aufgezeigt werden. Die deutsche Abstinenz an diesen Programmen war mit dem damaligen Bildungsbegriff begründet, nach welchem Bildung nicht messbar sei und die Leistungsstudien somit auch keine Defizite aufdecken könnten (vgl. Schwippert, 2005b, S. 3). Es bestand daher kein primäres Interesse an der empirischen Überprüfung von Schulqualität.

Diese „schwache Entwicklung einer empirisch ausgeprägten erziehungs-, bildungswissenschaftlichen und bildungspolitischen Denkweise“ (Hartung-Beck, 2009, S. 14) wandelte sich mit der einsetzenden intensiven Debatte über Evaluation und Optimierung von Schulqualität. In deren Konsequenz wurden seit Mitte der 1990er Jahre vermehrt bundeslandspezifische Leistungsmessungen durchgeführt. Zusätzlich nahm die Bundesrepublik 1995 erstmals

an einer internationalen Untersuchung, der TIMSS-Studie (Third International Mathematics and Science Study), teil. Diese Partizipation bewirkte eine reale empirische Wende in der Bildungspolitik und -forschung (vgl. Schwippert, 2005b, S. 3 f.). Die unterdurchschnittlichen mathematischen und naturwissenschaftlichen Leistungen der deutschen Schüler bei TIMSS erfuhren eine hohe Resonanz in der erziehungswissenschaftlichen und bildungspolitischen Debatte, so dass sich die Auseinandersetzung mit der Leistungsfähigkeit des Bildungssystems nochmals intensivierte. Als Reaktion hierauf legte die Ständige Konferenz der Kultusminister der Länder (KMK) mit den Konstanzer Beschlüssen 1997 die fortan regelmäßige Teilnahme an internationalen Schulleistungsstudien fest, was mit der Zielsetzung verbunden war, Optimierungs- und Steuerungsbedarf in den deutschen Bildungssystemen aufzudecken (vgl. Kultusministerkonferenz, 1997). Die bisherigen Annahmen über die Effektivität und Effizienz von Schule sollten durch eine empirische Datengrundlage ersetzt werden, welche zukünftig als Basis für bildungspolitische Reformen herangezogen werden sollte. Daher bescheinigt Steffens (vgl. 2007, S. 36) den TIMSS-Studien nachhaltige Auswirkungen auf die Systemsteuerung von Schule.

In Folge der empirischen Wende wurden weitere Leistungsvergleichsstudien auf regionaler und nationaler Ebene initiiert. Besonders bedeutend ist die Teilnahme Deutschlands an der PISA 2000-Untersuchung (Programme for International Student Assessment), deren Ergebnisse als sogenannter „PISA-Schock“ eine immense wissenschaftliche, politische sowie öffentliche Schuldebatte auslösten. Laut dieser Studie streuten in der Bundesrepublik die Schülerleistungen nicht nur stärker als im durchschnittlichen internationalen Vergleich, sondern es existierte außerdem eine verhältnismäßig große Risikogruppe an leistungsschwachen Schülern. Frappant waren zudem die Diskrepanzen der Schülerergebnisse zwischen den einzelnen Bundesländern und den verschiedenen Schulformen. Des Weiteren konnte ein signifikanter Zusammenhang zwischen der sozialen Herkunft, dem Migrationshintergrund und dem Bildungserfolg deutscher Schüler nachgewiesen werden (vgl. Drieschner, 2009, S. 22). Die Veröffentlichung der PISA-Ergebnisse stellte somit eine Zäsur in der Bildungspolitik dar und steigerte die intensive Diskussion um Schulqualität und eine notwendige Reform der Systemsteuerung.

Die KMK formulierte am 5./6. Dezember 2001 in Bonn - unmittelbar anknüpfend an die PISA-Ergebnisse - sieben Handlungsfelder, im Rahmen derer eine Bildungsreform angestoßen werden sollte. Für die Thematik dieser Arbeit ist das Handlungsfeld 5 von Bedeutung, welches „Maßnahmen zur konsequenten Weiterentwicklung und Sicherung der Qualität von Unterricht und Schule auf der Grundlage von verbindlichen Standards sowie eine ergebnisorientierte Evaluation“ forderte (Kultusministerkonferenz, 2001). Eine zukünftige

Konzentration auf die zentralen Ausrichtungen „Standardisierung“ und „Evaluation“ ist diesem Handlungsfeld bereits deutlich zu entnehmen. Spezifiziert wurde diese Orientierung mit dem Grundsatzbeschluss der KMK vom 23./24. Mai 2002, gemeinsame Standards für ausgewählte Schnittstellen als ein verbindliches Steuerungsinstrument für die allgemeinbildenden Schulformen zu erarbeiten (vgl. Kultusministerkonferenz, 2002a). Mithilfe der Standards sollen die verbindlichen Ziele des schulischen Lernens präzisiert werden. Die Einhaltung und Überprüfung der Standards werden über Orientierungs- und Vergleichsarbeiten sowie über eine wissenschaftliche Einrichtung vorgenommen. Dieser Form von externer Evaluation steht eine größere schulische Eigenverantwortung gegenüber, welche einer nachhaltigen Qualitätssicherung dient (vgl. Kultusministerkonferenz, 2002b). Begründet wurden die Reformbestrebungen im Sinne der Eigenverantwortung von Schulen zusätzlich mit der unterdurchschnittlichen Platzierung Deutschlands im internationalen Vergleich im Bereich der Autonomie von Schulen. Bei einem durchschnittlichen Mittelwert der Mitgliedsstaaten der OECD von 5,0 erreichte Deutschland lediglich einen Wert von 3,9 und befand sich somit im unteren Drittel (vgl. Avenarius, et al., 2003, S. 157 f.).

Das Konzept der eigenverantwortlichen Schule umfasst nicht die Loslösung der Einzelschule von der Bildungsadministration, sondern eine Verlagerung der Entscheidungskompetenzen von der Systemebene in die Einzelschule hinein, so dass dieses Prinzip als „Deregulierung“ bzw. als „Dezentralisierung“ bezeichnet werden darf (vgl. Altrichter & Rürup, 2010, S. 111 f.; Steffens, 2007, S. 33). Die strategischen Entscheidungsbefugnisse verbleiben in den höheren Systemebenen, während den Einzelschulen verstärkt operative Entscheidungsrechte und -kompetenzen übertragen werden (vgl. Altrichter & Rürup, 2010, S. 114). Mögliche Bereiche der Selbstständigkeit sind das Personalmanagement, die Ressourcenbewirtschaftung, die Konzeption eines Schulprofils und die Festlegung der Rahmenbedingungen des Unterrichts (vgl. Meetz, 2007, S. 119).

Die Eigenverantwortung impliziert eine vergrößerte pädagogische Freiheit, welche für die Sicherung und Weiterentwicklung der Schul- und Unterrichtsqualität genutzt werden kann. Hinter dieser Betrachtung steht die Intention, dass nur die Schule selbst die ihnen auferlegten Rahmenbedingungen bestmöglich an ihre individuellen schulischen, klassen- und schülerspezifischen Voraussetzungen anpassen kann. Die Einzelschule wird somit zum Zentrum der Qualitätsentwicklung (vgl. Drieschner, 2009, S. 30). Hierzu müssen Lehrpersonen und die Schulleitung über die dafür notwendigen Kompetenzen verfügen, um mit der hinzugewonnen Freiheit nachhaltig qualitätssichernd umzugehen.

Mit der Entwicklung zu einem neuartigen Verhältnis zwischen System- und Einzelschulebene bleibt der öffentliche Bildungsauftrag bestehen. Jedoch wird der Bedarf an der Festle-

gung von einheitlichen Zielerwartungen sichtbar. Die verstärkten Entscheidungs- und Gestaltungsrechte der Schulen sind aus diesem Grund mit der Verantwortung verknüpft, dass die Schüler die Zielvorgaben tatsächlich erreichen. Die Schulen sind daher mit einer Rechenschaftspflicht über die Resultate ihrer Lernprozesse konfrontiert, was wiederum eine „Rezentralisierung" bedeutet (vgl. ebd., S. 31). Die Rechenschaftslegung ist mit einer Doppelstrategie behaftet, bei der die Schule einerseits schulinterne Maßnahmen und Evaluationen durchführen muss, zum Beispiel mittels der gesetzlichen Verpflichtung zur Erstellung eines Schulprogramms. Andererseits werden über Vergleichsarbeiten und Schulinspektionsberichte die Lernresultate und Qualitätsentwicklungsprozesse verfolgt sowie auf die Bildungsstandards bezogen überprüft (vgl. Böttcher, 2009, S. 676; Steffens, 2007, S. 43).

Ob das Konzept der eigenständigen Schule kombiniert mit einer Rechenschaftspflicht tatsächlich eine Verbesserung der Lernergebnisse bewirkt, ist bislang in der empirischen Forschung nicht eindeutig belegt. Vielmehr wird von einer indirekten Wirkung ausgegangen, da die eigenverantwortliche Schule ihre Lernumgebung gezielter gestalten kann (vgl. Maag Merki & Steinert, 2006, S. 104; Meetz, 2007, S. 119 f.).

Verbunden mit dem Konzept der Eigenverantwortlichkeit ist die grundlegende Reform der Systemsteuerung, bei der die Wirkungen von Schule in das Zentrum der Diskussionen rücken (vgl. Höfer, Steffens, Diehl, Loleit, & Maier, 2009, S. 5). Die Steuerung des Bildungssystems und damit auch die Sicherung der Schulqualität waren zuvor durch eine Ausrichtung am Input determiniert gewesen. Für das Verständnis ist hierfür das „Modell zu Qualität und Qualitätssicherung im Bildungsbereich" von Ditton (2000a; 2000b) grundlegend, welches verkürzt betrachtet aus den drei Komponenten Voraussetzungen (Input), Prozesse und Ergebnisse (Output) besteht. Nach Drieschner wird unter der Inputsteuerung nun die „staatliche planerische Lenkung des Bildungswesens durch finanzielle und personelle Ressourcenzuweisung, rechtliche Vorschriften und inhaltliche Programmvorgaben" (2009, S. 25) verstanden. Für die Unterrichtspraxis bedeutete dies zentrale Vorgaben in Form von Lehrplänen und Unterrichtswerken, in denen die zu behandelnden inhaltlichen Themen detailliert aufgeschlüsselt sowie sortiert nach Jahrgangsstufen und den verschiedenen Schulformen zu entnehmen waren. Der Inputsteuerung lag zudem eine theoretische Annahme bezüglich der benötigten Lernzeit zugrunde, nach welcher alle Schüler den gleichen Unterrichtsstoff auf die gleiche Weise und in der gleichen Zeit zu lernen hätten. Den Lehrpersonen kam die elementare Aufgabe der Transformierung vorgegebener Inhaltsbereiche in didaktische Lernprozesse zu. Zudem wurde davon ausgegangen, dass für ein erfolgreiches Lernen vor allem das Angebot von Lerngelegenheiten entscheidend sei. Eine hochwer-

tige Lehrerausbildung und eine umfassende Ausstattung der Einzelschule seien als bedingende Faktoren anzusehen (vgl. Schwippert, 2005b, S. 5).

In der schulischen Praxis hat sich jedoch erwiesen, dass allein durch inhaltliche Vorgaben die Unterrichtsgestaltung und damit auch die Sicherung von Bildungsqualität nur schwer gesteuert werden können. Dieses traditionelle Instrument kann somit nicht den einzigen Baustein des Steuerungsmechanismus darstellen (vgl. Steffens, 2007, S. 42). In diesem Sinne äußerte sich 2005 die KMK nach vergleichender Beurteilung mit ausländischen Schulsystemen dahingehend, dass „die in Deutschland vorrangige Inputsteuerung allein nicht zu den erwünschten Ergebnissen im Bildungssystem führt. Die Festlegung und Überprüfung der erwarteten Leistungen müssen hinzu kommen“ (Kultusministerkonferenz, 2005a, S. 5).

Mit der Entwicklung der Bildungsstandards begann eine neuartige Orientierung am sogenannten Output, verbunden mit einer zunehmenden Distanzierung von der Inputsteuerung. Beim Output werden die Resultate schulischen Lernens, gemessen an den Schülerergebnissen, als Grundlage für eine Optimierung und systematische Evaluation der Lehr- und Lernprozesse herangezogen (vgl. Drieschner, 2009, S. 26 f.). Demnach wird der Output nach Klieme, et al. (vgl. 2007, S. 12) zukünftig zum maßgeblichen Bezugspunkt für die Beurteilung des Schulsystems in Hinblick auf seine Effizienz und Effektivität, indem durch systematische und regelmäßige Überprüfung der erreichten Leistungen der Lernenden Steuerungswissen zur Verfügung gestellt wird.

Die Systemsteuerung wird damit aus einem neuen Blickwinkel heraus vollzogen, da über die Analyse der (erwünschten) schulischen Ergebnisse auch Rückschlüsse auf den Input und die innerschulischen Prozesse gezogen werden können (vgl. Drieschner, 2009, S. 27) (vgl. Abbildung 1).

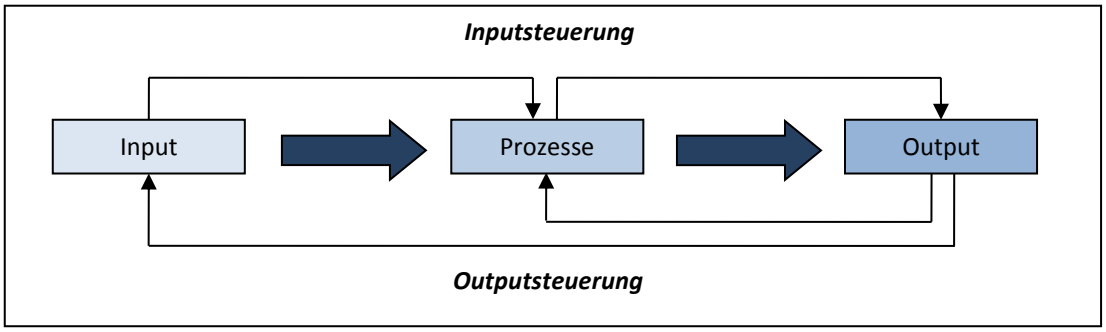

Abbildung 1: Vereinfachtes Modell der Input- und Outputsteuerung

Damit löst die Outputsteuerung nicht die Inputsteuerung ab. Vielmehr sind beide Elemente des Qualitätsmanagements auf Systemebene. Als weitere Steuerungsinstrumente können die zentralen Abschlussarbeiten sowie die Aufnahmeprüfungen für weiterführende Bil-

dungsgänge benannt werden, so dass in der Bundesrepublik eine Mischform zwischen verschiedenen Regulierungsmodellen vorliegt (vgl. Granzer, 2008, S. 50 f.). Grundsätzlich ist jedoch eine eindeutige Schwerpunktsetzung auf die Dimension des Outputs zu verfolgen. Daher kann in der Tat von einem Paradigmenwechsel hin zu einer outputorientierten Steuerung gesprochen werden.

Die wissenschaftliche Grundlage für die darauf folgende Konzeption sogenannter Bildungsstandards bildete die unter der Leitung von Klieme erstellte und 2003 veröffentlichte Expertise „Zur Entwicklung nationaler Bildungsstandards“ (Klieme, et al., 2007). Die Bundesrepublik folgt damit einem Trend, welcher seit Jahren bereits in den USA sowie in einer Vielzahl europäischer Staaten zu beobachten ist. Gerade in denjenigen Bildungssystemen der Länder, welche in der PISA-Untersuchung mit höheren Scores abgeschnitten hatten, sind Standards und Kernlehrpläne als Grundlage für das Bildungsmonitoring etabliert und dort werden zudem vermehrt Tests zur Gewinnung von Steuerungswissen eingesetzt (vgl. Klieme, 2003, S. 10; Steffens, 2007, S. 43). Die Bundesrepublik folgt dem Modell dieser Staaten in Hinblick auf eine systematische Qualitätssicherung mit zugehöriger Rechenschaftslegung.

Nachdem nationale Bildungsstandards in den folgenden Jahren entwickelt und verabschiedet wurden, formulierte die KMK in den Plöner Beschlüssen Juni 2006 ihre Gesamtkonzeption für die Maßnahmen zur Feststellung der Leistungsfähigkeit des Bildungssystems und der Schulen. Diese umfasst einerseits die weitere Teilnahme an den internationalen Leistungsvergleichsstudien PISA, TIMSS und IGLU (Internationale Grundschul-Lese-Untersuchung). Andererseits erfolgte eine Einigung auf die Durchführung von landesweit repräsentativen Stichprobenstudien zur Überprüfung der Standards. Diese Tests in den Domänen Sprachen, Mathematik und Naturwissenschaften werden in einem sechsjährigen Rhythmus und gekoppelt an den internationalen Untersuchungen durchgeführt. Zudem erfolgt zur Ermittlung von Lernständen auf Schulebene die Einführung von flächendeckenden Vergleichsarbeiten auf der Grundlage der Bildungsstandards (vgl. Kultusministerkonferenz, 2006). Mit diesem Maßnahmenkatalog wird die Reduzierung der durch PISA festgestellten Missstände im deutschen Bildungswesen angestrebt. Konkret bedeutet dies zum Beispiel die Verbesserung der Schülerleistungen und eine stärkere Fokussierung auf Fördermaßnahmen im Unterricht. Der zugrundeliegenden Annahme nach können nur dann die Bildungsqualität und das Bildungssystem optimiert werden, wenn die Zielstellungen hinreichend getestet und über die Resultate Informationen über den Erfolg von Lernprozessen generiert und genutzt werden (vgl. Maier, 2010b, S. 47). Gleichzeitig werden mittels der empirischen Studien und der Bildungsstandards die Leistungserwartun-

gen für die schulischen Akteure transparent, was sowohl der Einheitlichkeit als auch der Vergleichbarkeit der Schulabschlüsse über die Bundesländergrenzen hinweg dient.

2.2 Bildungsstandards

Im vorherigen Abschnitt wurde dargelegt, dass als zentrale Konsequenz aus den PISA-Ergebnissen heraus nationale Bildungsstandards von der KMK beschlossen wurden. Es ist für das Verständnis erforderlich, die Begrifflichkeit „Bildungsstandards“ sowie deren Funktionen weiterführend zu erläutern.

2.2.1 Begriffsbestimmung: „Bildungsstandards“

Beim Zugrundelegen einer humanistischen Vorstellung des Begriffes „Bildung“ werden Assoziationen mit einem lebenslangen Lern- und Entwicklungsprozess des Individuums verbunden. Unter dem Terminus „Standard“ ist hingegen eine gesetzte Norm zu verstehen. Mit den Bildungsstandards ist jedoch keineswegs eine Standardisierung oder Normierung der Persönlichkeitsentwicklung intendiert, wie eine naive Konklusion der beiden Begriffe vermuten lassen würde. Der Terminus „Bildungsstandards“ wurde in der deutschsprachigen erziehungswissenschaftlichen Forschung durch die Expertise von Klieme, et al. (2007) folgendermaßen nachhaltig geprägt.

> *Definition des Terminus „Bildungsstandards“ nach Klieme, et al.*
> „Bildungsstandards formulieren Anforderungen an das Lehren und Lernen in der Schule. Sie benennen Ziele für die pädagogische Arbeit, ausgedrückt als erwünschte Lernergebnisse der Schülerinnen und Schüler. Damit konkretisieren Standards den Bildungsauftrag, den allgemein bildende Schulen zu erfüllen haben.
> Bildungsstandards […] greifen allgemeine Bildungsziele auf. Sie benennen die Kompetenzen, welche die Schule ihren Schülerinnen und Schülern vermitteln muss, damit bestimmte zentrale Bildungsziele erreicht werden. Die Bildungsstandards legen fest, welche Kompetenzen die Kinder oder Jugendlichen bis zu einer bestimmten Jahrgangsstufe erworben haben sollen.“ (Klieme, et al., 2007, S. 19)

Bildungsstandards stellen folglich eine Vereinheitlichung der Basisqualifikationen dar, welche innerhalb der schulischen Bildungsbiographie gesichert werden müssen. Zudem können sie nach Köller (vgl. 2007, S. 97) als grundlegende Voraussetzungen für den erfolgreichen Erwerb vertiefter Allgemeinbildung – und somit in Richtung eines humanistischen Verständnisses von Bildung tendierend – betrachtet werden.

Wie aus der Definition hervorgeht, lehnen sich die Bildungsstandards an die Bildungsziele an, welche bereits 1973 von der KMK definiert wurden; sie sind jedoch keinesfalls mit ihnen gleichzusetzen. Während die Bildungsziele relativ allgemein gehaltene Aussagen über zu vermittelnde Fähigkeiten, Einstellungen und Werthaltungen formulieren, sind die Bildungsstandards von einer präzisen und fokussierten Ausrichtung an die zu erreichenden Lernergebnisse eines spezifischen Unterrichtsfaches gekennzeichnet (vgl. Klieme, et al., 2007, S. 20).

Weiterhin ist eine Abgrenzung der Standards zu dem anglikanischen Grundbildungskonzept „Literacy“ vorzunehmen. Standards postulieren definitiv nicht den Anspruch auf eine umfassende Abbildung von Allgemeinbildung, sondern sie orientieren sich stark funktional an den Grundprinzipien der Unterrichtsfächer (vgl. Köller, 2007, S. 97). Zusammenfassend charakterisiert Klieme (vgl. 2004, S. 10) die Bildungsstandards als einen Kompromiss zwischen den Fachdisziplinen, den Anforderungen der Lebens- und Arbeitswelt sowie den Entwicklungsbedürfnissen der Heranwachsenden.

2.2.2 Anforderungen an Bildungsstandards

Bei einer Analyse der etablierten Standard-Konzepte in den Bildungssystemen anderer Staaten können drei Typen von Standards klassifiziert werden:

Zum einen existiert die Form inhaltlicher Bildungsstandards (*Content Standards*), in welchen fachbezogene Leistungserwartungen hinsichtlich der verschiedenen Themengebiete ausdifferenziert werden. Ein weiterer Typus sind die Unterrichtstandards, auch als *Opportunity-to-learn-Standards* bezeichnet. In diesen prozessbezogenen Bildungsstandards werden Visionen von einem „guten“ und gelingenden Unterricht beschrieben. In Leistungsstandards (*Performance Standards*) werden hingegen allgemeine Fachkompetenzen als erwünschte Lernergebnisse formuliert, welche in verschiedenen inhaltlichen Dimensionen erworben werden können (vgl. Köller, 2007, S. 95). Performance Standards konzentrieren sich in besonderer Weise auf die erreichten Leistungen am Ende eines Lernabschnitts, so dass hiermit eine Überprüfung der Zielerwartungen verbunden ist. Aus diesem Grund eignet sich

dieser Standard-Typ insbesondere für outputorientierte Steuerungssysteme, welche eine Rechenschaftspflicht der Schulleistungen einfordern.

Zudem lässt sich eine weitere Klassifikation der Standards in Mindest-, Regel- und Maximalstandards vornehmen:

Mit den *Mindeststandards* wird ein basales Erwartungsniveau definiert, welches von allen Schülern, unabhängig von der jeweiligen Schulform, erreicht werden muss. Sie beschreiben somit eine zu erbringende Leistung, die nicht unterschritten werden darf und der verbindlichen Sicherung der Grundbildung dient (vgl. Drieschner, 2009, S. 29, 58). Auf diese Weise wird insbesondere von den leistungsschwachen Schülern ein Grundniveau eingefordert. Der Befürchtung von Criblez, et al. (vgl. 2009, S. 29), welche mit den Mindeststandards die Gefahr einer Minimalisierung des Lehrens und Lernens assoziieren, ist entgegenzuhalten, dass gerade auf Basis der Mindeststandards bundeslandweit oder durch die Einzelschule selbst weitere Niveaustufen entwickelt werden können, welche höhere Anforderungen beinhalten. Die Entwicklung von validen Mindeststandards, die tatsächlich von allen Schülern erreicht werden, setzt allerdings einen umfangreichen empirischen Kenntnisstand über die derzeitigen Kompetenzverteilungen der Schüler voraus, welcher zuvor erhoben werden müsste (vgl. Artelt & Riecke-Baulecke, 2004, S. 21). Des Weiteren muss geklärt sein, welche Maßnahmen für den Fall zu ergreifen sind, dass Schüler die Mindeststandards und somit das Bildungsziel verfehlen.

Als weitere Form sind die *Regelstandards* zu benennen, welche ein durchschnittliches Erwartungsniveau ausweisen. Ausgehend von der Annahme einer Normalverteilung der Schülerleistungen sind diese Standards in der Form konzipiert, dass die Mehrheit der Schüler die Anforderungen erfüllt und ein gewisser Anteil bessere bzw. schlechtere Leistungen aufweist. Dieser Standard-Typ würde den deutschen Schulsystemen die erwiesenen Charakteristika der massiven Selektion und der großen Zahl an scheiternden Schülern ohne Schulabschluss erneut manifestieren. Problematisch ist vor allem, dass diese Standards nicht ausdrücken können, was ein Schüler können muss, um als ein erfolgreich Lernender zu gelten (vgl. Klieme, et al., 2007, S. 28). Daher könnten nach Criblez, et al. (vgl. 2009, S. 30) bei der Leistungsbewertung keine kriterialen Bezugsnormen herangezogen werden, sondern lediglich soziale (zu den Bezugsnormen vgl. Abschnitt 4.3.6.2). Lersch (vgl. 2006, S. 31) kritisiert weiterhin am Konzept der Regelstandards, dass deren Formulierung im Gegensatz zu klaren und verbindlichen Mindeststandards eher weich und unscharf ausfallen würde. Zudem käme es zu Differenzierungsproblemen in Hinblick auf die Förderung von Schülern, welche sich sowohl unterhalb als auch oberhalb der gesetzten Standards befänden.

Eine nochmalige Steigerung in der Ausweisung von zu erwartenden Leistungsergebnissen wird in Form von *Maximalstandards* vorgenommen. Diese definieren ein Höchstniveau an zu erreichenden Kompetenzen, welche sich nicht mehr an den durchschnittlichen realen Schülerleistungen orientieren (vgl. Ziener, 2006, S. 50 f.). Weil nur eine Minderheit von etwa 10 Prozent diese Standards erreichen kann, ergibt sich eine andauernde Überforderung der Schüler im Unterricht. Als eine daraus resultierende negative Konsequenz ist das Förderproblem zu nennen, denn 90 Prozent der Lernenden wären nach diesem Konzept als förderbedürftig einzuordnen (vgl. Lersch, 2006, S. 29). Ebenso ist der demotivierende Aspekt dieser Standards nicht zu unterschätzen, dass sie mit einer Defizitorientierung einhergehen, da die erbrachte Schülerleistung stets mit der negativen Abweichung von den Maximalstandards angegeben wird (vgl. Klieme, et al., 2007, S. 28). Dennoch kommt diesem Typ von Bildungsstandards eine Entwicklungsfunktion zu, indem mit ihnen die Richtung, in welcher sich die Bildungsqualität entwickeln sollte, beschrieben werden kann (vgl. Zeitler, Köller, & Tesch, 2010, S. 30).

In der Expertise von Klieme, et al. (vgl. 2007, S. 24 ff.) wurde ein an die Schulsysteme Deutschlands angepasstes Konzept der Standardisierung entwickelt, welches die Grundlage für die darauf folgende Konstruktion der Bildungsstandards darstellte. Hierbei wurden sieben zentrale Anforderungen an die Bildungsstandards und deren Umsetzung benannt:

1. *Fachlichkeit* in Bezug auf bestimmte Lernbereiche, verbunden mit der Hervorhebung der Grundprinzipien eines Faches. Hierbei sollten auch fächerübergreifende Bildungsziele berücksichtigt werden.

2. *Fokussierung* auf den Kernbereich eines Faches, was dem Anspruch der traditionellen Lehrpläne, die gesamte Breite einer Fachdisziplin abzubilden, entgegensteht.

3. *Kumulativität* des systematischen Kompetenzerwerbs als Resultat des Lehrens und Lernens im Unterricht. Diesem Prinzip liegt die Vorstellung eines nachhaltigen und aufeinander aufbauenden Lernens zugrunde.

4. *Verbindliche Mindeststandards*, die für alle Schüler schulformübergreifend gelten. Hiermit wird die Zielstellung verfolgt, insbesondere die Leistungsschwachen nicht zurückzulassen.

5. *Differenzierung* durch die Entwicklung von Kompetenzmodellen, in welchen die Kompetenzen hinsichtlich ihrer Dimensionen und Anforderungsstufen weiter ausdifferenziert werden. Ziel ist ein für alle Schulformen gleichermaßen geltender Kompetenzrahmen, aus welchem im zweiten Schritt unterschiedliche Niveaus abgeleitet werden

können. Auf dieser Basis können Einzelschulen weiterführend ein schuleigenes Curriculum entwickeln bzw. einer spezifischen Profilrichtung folgen.

6. *Verständlichkeit* durch eine klare, knappe und nachvollziehbare Formulierung der Bildungsstandards.

7. *Realisierbarkeit und Überprüfbarkeit*, indem die Standards einerseits unterrichtbar sein müssen und die Umsetzung im Unterricht durch exemplarische Operationalisierungen gefördert werden sollte. Andererseits müssen die zu erwartenden Lernergebnisse so klar ausgedrückt sein, dass valide Überprüfungen in Form von Tests möglich sind.

Köller (vgl. 2010, S. 531) führt zusätzlich die Anforderungen *Abschlussbezug* und *länderübergreifende Gültigkeit* an. Um den genannten Anforderungen zu genügen, müssen bei der Entwicklung von Standards sowohl bildungspolitische als auch pädagogische und fachdidaktische Grundprinzipien berücksichtigt werden.

2.2.3 Funktionen der Bildungsstandards

Da mit den Bildungsstandards vielfältige Funktionen verbunden werden, ist ihre Einordnung in die drei Ebenen Makroebene (Systemebene), Mesoebene (Einzelschulebene) und Mikroebene (Unterrichtsebene) sinnvoll.

Makroebene (Bildungssystem)

Die Bildungsstandards üben als neues Steuerungsinstrument zwei wesentliche Funktionen auf der Systemebene aus: Zum einen dienen sie über die Formulierung verbindlicher Leistungserwartungen der Qualitätssteigerung des Bildungssystems. Zum anderen nehmen sie eine Überprüfungsfunktion wahr, indem die erreichten Schülerleistungen anhand der zu erwerbenden Kompetenzen gemessen werden. Auf diese Weise kann sowohl eine Qualitätssicherung als auch ein Bildungsmonitoring stattfinden, auf dessen Grundlage wiederum Verbesserungs- bzw. Optimierungsbedarf abgeleitet werden kann (vgl. Köller, 2007, S. 95; Lersch, 2006, S. 31). Zudem wird mit den Bildungsstandards die Erwartung verknüpft, die bundesweite Vergleichbarkeit und Anschlussfähigkeit der verschiedenen Schulabschlüsse zu stärken. Einher geht damit die Zielstellung einer erhöhten Durchlässigkeit zwischen den einzelnen Schulformen (vgl. Artelt & Riecke-Baulecke, 2004, S. 7; Drieschner, 2009, S. 24). Dies kann jedoch nur mithilfe der Konzeption von Mindeststandards gewährleistet werden, welche national und schulformübergreifend für alle Schüler gleichermaßen verbindlich sind.

Mesoebene (Einzelschule)

Mithilfe der Bildungsstandards kann das interne Evaluationsprogramm einer Schule maßgeblich verbessert werden, da sie klare Kriterien und Vergleichsmaßstäbe bieten (vgl. Demmer & Schweitzer, 2005, S. 69). Beispielsweise ist die Formulierung wesentlicher und prägnanter Unterrichtsziele im Schulprogramm ein Schritt in diese Richtung. Ebenso kann ein auf den Bildungsstandards fußendes schulinternes Lerncurriculum entwickelt und erprobt werden. Möglich wird dies über Freiräume für die innerschulische Lehr- und Lernplanung, die durch die Bildungsstandards zur Verfügung gestellt werden. Somit passt sich das Steuerungsinstrument „Bildungsstandards" zugleich in das Prinzip der eigenverantwortlichen Schule ein. Verbindliche Zielvorgaben des Lernens werden zwar vorgegeben, die Schule entscheidet jedoch selbst, auf welchem Weg die Kompetenzen von den Schülern erworben werden.

Mikroebene (Unterricht)

Als weitere Funktion der Bildungsstandards ist der Anstoß von Unterrichtsentwicklung anzuführen, so dass mit den Standards direkte Konsequenzen für die Durchführung von Lernsituationen intendiert werden sollen. Die Bildungsstandards dienen der Lehrperson als Orientierung hinsichtlich der Planung, der Analyse und der Überprüfung des eigenen Unterrichts, da sie ein einheitliches Referenzsystem für das pädagogische Handeln vorgeben (vgl. Klieme, et al., 2007, S. 9). Die Analyse des Leistungsstandes der einzelnen Schüler wird durch zentrale Leistungsfeststellungen erheblich erleichtert. Auf diese Weise kann eine frühzeitige Diagnose von Förderbedarf erfolgen. Die Standards sollen dabei die Berücksichtigung unterschiedlicher Lernausgangslagen im Unterricht befördern, wie zum Beispiel durch die zunehmende Verwendung von Methoden, bei welchen die Schüler selbstverantwortlich und selbstgesteuert lernen. Dies trägt zusätzlich zu einer Stärkung des kompetenten Umgangs mit Heterogenität bei der Lehrkraft bei. Darüber hinaus sollen die Bildungsstandards dazu genutzt werden, den eigenen Unterricht kritisch zu hinterfragen und die Diagnosefähigkeit zu trainieren. Für die Schüler sowie deren Eltern werden die an sie gestellten Leistungserwartungen durch die Standards transparenter (vgl. Granzer, 2008, S. 55).

Letztendlich liefern die Bildungsstandards einen wesentlichen Beitrag zur Reduzierung der in PISA etc. festgestellten Missstände im Bildungswesen. Sie folgen jeweils den Funktionen der *Orientierung* und der *Steuerung*. Wesentlich ist hierbei, dass die Standards keineswegs als Selektions- oder Kontrollinstrument verwendet oder missbraucht werden dürfen. Vielmehr sollten sie als ein zentraler Bestandteil des umfassenden Programms zur Entwicklung

und Sicherung der Bildungs-, Schul- und Unterrichtsqualität interpretiert werden (vgl. Kultusministerkonferenz, 2005a, S. 10).

2.2.4 Konzeption und Implementierung der nationalen Bildungsstandards

Die KMK übernahm die Aufgabe, nationale Bildungsstandards festzulegen, welche für alle Bundesländer verpflichtend implementiert werden. Die Konzeptionalisierung der Standards wurde in einer interdisziplinären Kooperation von Bildungswissenschaftlern, Fachdidaktikern und Lehrkräften vorgenommen (vgl. Köller, 2007, S. 97). Am 4. Dezember 2003 erfolgte zunächst in Bonn der Beschluss über die Standards für den Mittleren Schulabschluss in den Fächern Deutsch, Mathematik und der ersten Fremdsprache (Englisch/ Französisch) (vgl. Kultusministerkonferenz, 2003). Erweitert wurde diese Grundlage am 14./ 15. Oktober 2004 in Mettlach durch Primarstufenstandards in Deutsch und Mathematik sowie durch Bildungsstandards für den Hauptschulabschluss in Deutsch, Mathematik und der ersten Fremdsprache (vgl. Kultusministerkonferenz, 2004a; 2004c). Am 16. Dezember 2004 beschloss die KMK in Bonn zusätzlich, auf den Mittleren Schulabschluss bezogene Standards für die naturwissenschaftlichen Fächer Chemie, Biologie und Physik zu verabschieden (vgl. Kultusministerkonferenz, 2004b).

Die Bildungsstandards in den Hauptfächern für den Mittleren Schulabschluss wurden bereits im Schuljahr 2004/2005 implementiert und sollen im Unterricht als Grundlage für die Leistungsanforderungen Anwendung finden. Die übrigen Bildungsstandards gelten seit dem Schuljahr 2005/2006 als verbindlich. 2007 folgte des Weiteren der Beschluss, auch für die Allgemeine Hochschulreife Standards zu konzipieren (vgl. Kultusministerkonferenz, 2007).

Die Auswahl der Fächer lässt eine spürbare Konzentration auf die Hauptfächer erkennen. Dies liegt laut Aussage der KMK darin begründet, dass in Deutsch, Mathematik und der ersten Fremdsprache die wesentlichen Fähig- und Fertigkeiten erlangt würden, die für die spätere Weiterbildung und im Alltag notwendig seien. Die Naturwissenschaften wurden mit Standards versehen, weil diese Fächer in PISA und TIMSS ebenfalls Untersuchungsgegenstände sind (vgl. Kultusministerkonferenz, 2005a, S. 14).

Des Weiteren ist als Merkmal der nationalen Bildungsstandards festzustellen, dass sie *abschlussbezogen* konzipiert worden sind. Die Abschlüsse werden in diesem Zusammenhang als Schnittstellen im Bildungsweg eines Menschen interpretiert. Zusätzlich wird das Ziel verfolgt, mit einer solchen Ausrichtung eine Vergleichbarkeit der Schulabschlüsse zu gewährleisten (vgl. Thies, 2005, S. 12). In Bezug auf die möglichen Kategorisierungen von Standards lassen sich die nationalen Bildungsstandards als *Performance Standards* charak-

terisieren, die angeben, welche Kompetenzen die Schüler am Ende eines Bildungsabschnitts erreicht haben sollen. Entgegen der Forderung in der Expertise von Klieme, et al. (vgl. 2007, S. 9) nach Mindeststandards entsprechen die von der KMK beschlossenen Standards dem Typus der *Regelstandards*. Dies lässt sich mit dem Zeitdruck begründen, in welchem der gesamte Reformprozess verlief. Wie bereits erläutert, wird für die Konzeption von Mindeststandards eine umfangreiche empirische Datenanalyse benötigt, um darauf aufbauend die Kompetenzverteilungen zu ermitteln. Diese lagen bei der Konstruktion der nationalen Bildungsstandards noch nicht vor, so dass vorläufig Regelstandards verabschiedet wurden, um eine Unter- oder Überforderung der Schüler zu vermeiden (vgl. Artelt & Riecke-Baulecke, 2004, S. 21). Mittlerweile wurden die ersten Kompetenzstufenmodelle konzipiert. Es bleibt abzuwarten, ob die Regelstandards in Mindeststandards transformiert werden. Zumindest kann bereits anhand der Aufgabenbeispiele, die in der Stufenskala dem Kompetenzniveau II zugeordnet werden, abgelesen werden, welche Minimalanforderungen erwartet werden (bezüglich der Zuordnung von Mindest-/Regel-/Maximalstandards zu Kompetenzstufen anhand von Aufgaben vgl. Abschnitt 2.3.2.2).

In Bezug auf die an Bildungsstandards gestellten Anforderungen kann festgehalten werden, dass insbesondere die Standards für die Fächer Deutsch und Mathematik den Kriterien Verbindlichkeit und Differenzierung und teilweise auch der Kumulativität, Verständlichkeit, Realisierbarkeit und Messbarkeit nicht zufriedenstellend gerecht werden (vgl. Köller, 2010, S. 534; Zeitler, Köller, & Tesch, 2010, S. 24 ff.).

Die Kultusministerien der Bundesländer tragen die Verantwortung für die Verankerung der Bildungsstandards in der Schulpraxis. Um darüber hinaus dennoch einen nationalen Rahmen zu schaffen, wurde am 3./4. Dezember 2004 in Mainz von der KMK die Gründung eines Begleitinstituts beschlossen. Dieses Institut zur Qualitätsentwicklung im Bildungswesen (IQB) ist eine wissenschaftliche Einrichtung der Länder an der Humboldt-Universität zu Berlin e.V. Zu dessen Aufgaben gehören beispielsweise die Erstellung und Skalierung von Aufgabenpools, auf deren Basis Kompetenzstufenmodelle konstruiert werden (vgl. Hofmann-Göttig, Eschmann, & Daumen, 2005, S. 34). Ebenso soll eine Weiterentwicklung der Bildungsstandards forciert und wissenschaftlich ausgewertete Aufgabensammlungen angelegt werden. Das IQB verfolgt zudem das Ziel, die Reformbemühungen der Bundesländer im Kontext der Bildungsstandards zu stärken und einen länderübergreifenden Austausch zu fördern.

Die Festlegung der inhaltlichen Lernziele bleibt hingegen weiterhin den Bundesländern überlassen. Für eine stetige Konkretisierung der Bildungsstandards wurden in den Ländern Arbeitsgruppen bzw. Landesinstitute gegründet. Diese Einrichtungen sind den Kultusmini-

sterien unmittelbar nachgeordnete Behörden, die vorrangig eine Beratungs- und Unterstützungsfunktion ausüben (vgl. Meinel & Sachse, 2007, S. 261 ff.). Zum Beispiel geschieht dies durch eine enge Zusammenarbeit mit Fortbildungsstellen, der Herausgabe von Handreichungen und Materialien oder der Veranstaltung von Informationszirkeln. Des Weiteren sind die Landesinstitute für die Erstellung von Kerncurricula (vgl. Abschnitt 2.5) und der zentralen Abschlussarbeiten zuständig. Für den Kontext dieser Arbeit ist es besonders bedeutsam, dass die Landesinstitute Orientierungshilfen bzw. Zwischenstandards für weitere Klassenstufen entwickeln. Hiermit wird die Grundlage sowohl für die Umsetzung eines kumulativen kompetenzorientierten Unterrichts (vgl. Abschnitt 2.6) als auch für das zwischenzeitliche Überprüfen der Standards gelegt. Ebenso werden in einigen Bundesländern weitere Fächer mit Bildungsstandards versehen, die von der KMK nicht berücksichtigt wurden (vgl. Korngiebel, 2009, S. 22 f.).

Mithilfe der Arbeiten in den Landesinstituten wird nach Drieschner (vgl. 2009, S. 24) eine strukturelle und inhaltliche Annäherung der Bildungspläne der einzelnen Bundesländer gefördert. Hierfür erfolgt ein intensiver länderübergreifender Austausch der jeweiligen Institute in einem initiierten „Netzwerk für empiriegestützte Schulentwicklung“ (EMSE), welches durch wissenschaftliche Berater unterstützt wird.

2.3 Kompetenzen

2.3.1 Begriffsbestimmung: „Kompetenzen“

Die von der KMK entwickelten Bildungsstandards sind nicht wie die bisherigen Lehrpläne in Form von Inhaltsauflistungen gestaltet, sondern sie drücken Kompetenzen aus. Der Kompetenzbegriff stellt ein theoretisches sowie empirisch fundiertes Konstrukt der sozial- und erziehungswissenschaftlichen Forschung dar (vgl. Klieme & Hartig, 2007, S. 14). Dennoch existieren verschiedene Varianten und Auslegungen, welche von einem rein fachlichen Bezug bis zu einer starken Annäherung zum Intelligenzbegriff heranreichen (vgl. Hartig & Klieme, 2006, S. 128 f.). Weinert schlägt im Zusammenhang mit den Bildungsstandards folgende Definition von Kompetenz vor, welche wiederum in der Expertise von Klieme, et al. (2007) sowie in dem daraufhin eingesetzten erziehungswissenschaftlichen Diskurs als Arbeitsgrundlage verwendet wurde:

Definition Kompetenz nach Weinert

„[Kompetenzen sind] die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können." (Weinert, 2001, S. 27)

Kompetenzen sind im Unterschied zur Intelligenz erlernbare Verhaltens- und Leistungsdispositionen, welche kontext- und situationsbezogen verantwortungsvoll eingesetzt werden können. Damit ist der Anspruch verbunden, mithilfe von Kompetenzen spezifischen Handlungsanforderungen gerecht zu werden. Als pädagogisches Ziel des Kompetenzerwerbs ist nach Klieme & Hartig die „Befähigung zu selbstständigem und selbstverantwortlichem Handeln und damit zur Mündigkeit" (2007, S. 21) zu benennen.

Kompetenzen sind funktional, indem sie sich auf einen spezifischen Sektor von Kontexten oder Situationen beziehen. Daher sind sie durch ihre Domänenspezifität gekennzeichnet (vgl. Hartig & Klieme, 2006, S. 129). Aufgrund dieses Charakteristikums lässt sich nach der Definition von Weinert eine eindeutige Abgrenzung vom Konzept des Intelligenzbegriffs, welches allgemeine kognitive Grundfähigkeiten umfasst und somit domänenunabhängig ist, vornehmen (vgl. ebd., S. 130). Kompetenzen sind hingegen fest umrissen und überprüfbar. Auf diese Weise kann mithilfe spezieller Messinstrumente untersucht werden, welche Kompetenzen ein Schüler in welchem Ausmaß erworben hat.

Lersch (vgl. 2006, S. 33) und Klinger (vgl. 2005, S. 137) spezifizieren den Kompetenzbegriff um die zwei Dimensionen *Wissen* und *Können*. Während das Wissen aus kognitiven Strukturen, Einstellungen und Haltungen besteht, zeigt sich das Können in konkreten Situationen in Form von operationalen Aktivitäten. Beide Dimensionen sind notwendig für ein nachhaltiges Lernen im Sinne eines kumulativen Kompetenzerwerbs. Handeln ohne das nötige Hintergrundwissen, was und warum man etwas wie tut, ist wenig erfolgsfördernd. Andersherum können höhere Kompetenzanforderungen nicht ausschließlich mithilfe der Theorie bewältigt werden. Vielmehr ist eine zunehmende Prozeduralisierung stets erforderlich. Beide Komponenten bedingen sich somit gegenseitig und sind Voraussetzung für die Aneignung einer Kompetenz. Das wechselseitige Zusammenwirken von Fähigkeiten, Fertigkeiten und Einstellungen bildet letztlich die Grundlage für die Ausbildung von *Handlungskompetenzen*, unter welchen die Befähigung verstanden wird, sich in verschiedenen Lebenssituationen sachgerecht, durchdacht sowie individuell und sozial verantwortlich zu verhalten (vgl. Staatsinstitut für Schulqualität und Bildungsforschung München, 2006, S. 2). Hierfür ist zu-

nächst die Bereitstellung geeigneter Lehr- und Lernsituationen erforderlich, in denen der Schüler seine erworbene Kompetenz unter Beweis stellen kann (vgl. Lersch, 2006, S. 32 f.).
Des Weiteren kann eine Unterscheidung zwischen domänenspezifischen *Fachkompetenzen* und allgemeinen *überfachlichen Kompetenzen* vorgenommen werden. Bei den fachlichen Kompetenzen stehen der Erwerb inhaltlichen Wissens sowie Strategien zur praktischen Nutzung des Erlernten im Blickfeld (vgl. Lersch, 2007, S. 435). Fächerübergreifende Kompetenzen entwickeln sich dagegen auf Basis kognitiver, motivationaler, volitionaler und emotionaler Einstellungen und werden in verschiedenen Lerngebieten gefordert und gefördert (vgl. Klieme, Artelt, & Stanat, 2001, S. 204). Fächerübergreifende Kompetenzen sind nach dieser Begriffsbestimmung eher allgemeine und notwendige Fähigkeitsdimensionen, um das zu erwerbende Wissen und Können über einen strukturierten und effektiven Lernvorgang zu steuern. Ein Beispiel hierfür ist die Problemlösekompetenz für neuartige Anforderungssituationen. Diese Kompetenz wurde theoretisch sowie empirisch umfassend untersucht und bereits in PISA getestet. Andere überfachliche Kompetenzen wie zum Beispiel Lern-, Selbst-, Sozial- und Methodenkompetenzen sind teilweise relativ unspezifisch. Aufgrund der allgemeinen Formulierungen wird die Operationalisierung erschwert, was sich insbesondere beim Überprüfen und Testen der Kompetenzen negativ auswirkt. Im Unterricht werden die überfachlichen Kompetenzen mithilfe von besonderen Lerngelegenheiten im Kontext fachlicher Lehr-Lern-Prozesse erworben. Beispielsweise kann ein Schüler nicht das Lernen lernen, ohne zugleich fachliche Aspekte zu lernen. Daher wird ein kompetenzorientierter Unterricht Wissen und Können sowie fachliche und selbstregulative Kompetenzen über Lerntransfers auf verschiedenen Ebenen gleichermaßen berücksichtigen (vgl. Lersch, 2007, S. 438 f.). Die Grundzüge der kompetenzorientierten Unterrichtsdidaktik werden in Abschnitt 2.6 erläutert.
Ob ein erfolgreicher Kompetenzerwerb stattfindet, kann ausschließlich in situierten Handlungssituationen festgestellt werden, in denen sich die Kompetenz manifestiert. Bedeutsam ist hierbei sowohl die Art und Weise als auch der bereits erreichte Grad der Kompetenzausprägung. Zusammenfassend verfügen Schüler laut der KMK über eine Kompetenz, wenn sie

- „zur Bewältigung einer Situation vorhandene Fähigkeiten nutzen,
- dabei auf vorhandenes Wissen zurückgreifen und sich benötigtes Wissen beschaffen,
- die zentralen Zusammenhänge eines Lerngebietes verstanden haben,
- angemessene Lösungswege wählen,
- bei ihren Handlungen auf verfügbare Fertigkeiten zurückgreifen,

- ihre bisher gesammelten Erfahrungen in ihre Handlungen mit einbeziehen.“ (Kultusministerkonferenz, 2005a, S. 16)

Diese Kriterien folgen somit der erweiterten Definition von Weinert, indem fächerübergreifende Kompetenzen als notwendige Bestandteile eines nachhaltigen, lebenslangen Lernens betrachtet werden. Es ist hierbei anzumerken, dass in den nationalen Bildungsstandards der KMK lediglich fach- und domänenbezogene Kompetenzen ausgewiesen wurden. Fächerübergreifende Kompetenzen konstituieren somit keine eigenen Standards. Diese Konzeption folgt dem Verständnis, dass zunächst fachliche Kompetenzen als Grundlage für überfachliche Kompetenzen erworben werden. Die Bildungsstandards sind eng an die fachlichen Disziplinen gebunden, daher haben die Lehrkräfte eigenverantwortlich zu entscheiden, an welchem inhaltlichen Kontext überfachliches Lernen stattfinden kann. Die daraus resultierenden Probleme für eine Überprüfbarkeit der Standards in Form von Tests wurden bereits angesprochen. Aus diesem Grund wird in derzeitigen Forschungsaktivitäten, wie dem integrativen DFG-Schwerpunktprogramm unter Leitung von Klieme & Leutner (2006), in welchem Modelle von Kompetenzen für spätere Diagnostik- und Testverfahren empirisch entwickelt werden, eine eingeschränkte Version des Kompetenzbegriffs verwendet, indem zunächst nur auf den kognitiven Bereich und somit auf die fachlichen Kompetenzen Bezug genommen wird. Es erfolgt daher eine Abgrenzung von Handlungskompetenzen, die motivationale und affektive Einstellungen und Tendenzen einschließen würden.

2.3.2 Kompetenzmodelle

Zur weiteren Präzisierung der Bildungsstandards werden vom IQB Kompetenzmodelle entwickelt. Mithilfe dieser wissenschaftlichen Konstrukte können empirisch validierte Aussagen darüber getroffen werden, in welchen Kontexten und in welchen Lernentwicklungsphasen die Schüler welche Kompetenzen erwerben. Dies bietet die Möglichkeit, den erlangten Kompetenzgrad der Lernenden präzise zu bestimmen, so dass die Kompetenzmodelle für die spätere Überprüfung der Bildungsstandards und des Lernstandes ein zentrales Instrument darstellen. Die Modelle, welche interdisziplinär von Experten der Pädagogik, Psychologie und der Fachdidaktik entwickelt werden, bilden sowohl kumulative Lernprozesse in Form eines Längsschnitts als auch nicht-kumulative Elemente ab (vgl. Klieme & Leutner, 2006, S. 885). Hierzu werden zum einen die Kompetenzen in ihren inhaltlichen Domänen in Form von Kompetenzstrukturmodellen weiter ausdifferenziert. Zum anderen werden Ni-

veaus über Kompetenzstufenmodelle definiert, in denen der erreichte Lernstand verortet werden kann. Im Folgenden werden diese Modelle vertieft vorgestellt.

2.3.2.1 Kompetenzstrukturmodelle

Kompetenzstrukturmodelle – auch Komponentenmodelle genannt – beschreiben das inhaltliche Anforderungsgefüge der Kompetenzen. Durch die Bildung von Kompetenzbereichen werden ihre Dimensionen aufgezeigt. Die Entwicklung der Kompetenzstrukturmodelle wird mithilfe einer faktorenanalytischen Methode vorgenommen. Die Leitfrage ist hierbei, welche Fähig- und Fertigkeiten in bestimmten Zusammenhängen erfasst werden können. Kompetenzen, die eine hohe Korrelation zueinander aufweisen, beziehen sich nach der Interpretation dieser Methode auf dasselbe Merkmal (vgl. Hartig & Klieme, 2006, S. 132). Auf diese Weise können sie zu übergeordneten Bereichen zusammengefasst werden, so dass eine strukturierte und zugleich präzis differenzierte Formulierung der Kompetenzen ermöglicht wird.
Die Problematik der alleinigen Berücksichtigung von fachlichen Kompetenzen in den Bildungsstandards wird in den Kompetenzstrukturmodellen erneut ersichtlich. Eine eindeutige Trennung zwischen fachlichen und überfachlichen Kompetenzen scheint oftmals nicht möglich zu sein. Die Fähigkeit des Faches Deutsch „mit Texten umgehen“ korreliert beispielsweise stark mit Methodenkompetenzen. Dies trifft ebenso auf die erste Fremdsprache zu: interkulturelle Kompetenzen vereinen sowohl Fach-, als auch Selbst- und Sozialkompetenzen in sich. Aufgrund dieser Mehrdimensionalität ist es bei der Entwicklung von charakteristischen Aufgabenbeispielen und der späteren Messung erforderlich, die enthaltenen überfachlichen Komponenten nicht zu berücksichtigen.

2.3.2.2 Kompetenzstufenmodelle

Für einen kumulativen Kompetenzerwerbsprozess ist die Diagnostik des bereits erreichten Lernstandes von enormer Bedeutung. Es stellt sich die Frage, welche Anforderungen zum Beispiel ein Schüler mit einer stark bzw. schwach ausgebildeten Kompetenz bewältigen kann. Die quantitativ abgebildeten Leistungswerte in Form von Messwerten oder Schulnoten bedürfen demnach einer kriteriumsorientierten Interpretation, um einem Lernenden den erworben Kompetenzgrad zuweisen zu können. Aus diesem Grund werden Kompetenzstufen- bzw. Kompetenzniveaumodelle konstruiert, welche verschiedene Anforderungsstu-

fen empirisch validiert abbilden. Diese Modelle widersprechen dem System der herkömmlichen Lehrpläne, welche implizit die Forderung erheben, dass alle Schüler die Ziele im gleichen Maße zu erreichen haben. Mit der Entwicklung von Stufen werden hingegen ein realistischer Erwartungshorizont und die Voraussetzungen für einen auf Heterogenität ausgerichteten und differenzierenden Förderunterricht geschaffen.

Die Kompetenzstufenmodelle lösen in gewisser Weise die bisher verwendeten Anforderungsbereiche ab. Diese dienten bislang als Orientierungshilfe für die Lehrkräfte und untergliederten sich für die einzelnen Fächer in jeweils drei Bewältigungsstufen. Die Problematik lag an dieser Stelle bei den unspezifischen und allgemein gehaltenen Formulierungen, die für jeden Bildungsgang und jede Altersstufe gleichermaßen galten. Über die Qualität und die Komplexität der Anforderungsbereiche seien nach Äußerung der KMK keine wissenschaftlichen Aussagen möglich gewesen (vgl. Kultusministerkonferenz, 2005b, S. 17). Eine Schülerleistung konnte somit anhand der Anforderungsbereiche nicht differenziert genug erfasst werden. Erste Kompetenzstufenmodelle wurden daraufhin in TIMSS und PISA verwendet, bei denen die jeweilige Aufgabenschwierigkeit dem Konstrukt von fünf Stufen zugeordnet wurde. Zu beachten ist hierbei, dass die TIMSS- und PISA-Modelle lediglich die querschnittliche Leistungsverteilung der untersuchten Schülergruppen ermittelten und sie daher nicht als Lernentwicklungsmodelle zu bewerten sind (vgl. Klieme, et al., 2007, S. 76 f.).

Bei Kompetenzstufenmodellen handelt es sich folglich um eine Beschreibung, welche konkreten situativen Anforderungen eine Personen bei einer bestimmten Kompetenzausprägung bewältigen kann (vgl. Klieme & Leutner, 2006, S. 883). Hierfür wird eine kontinuierliche Kompetenzskala konstruiert, welche in fünf Abschnitte, den sogenannten Kompetenzstufen, unterteilt wird (vgl. Abbildung 2).

Jeder Kompetenzstufe wird ein dementsprechend hoher Kompetenzgrad zugeordnet. Auf diese Weise kann eine kriteriale Bewertung von gemessenen Leistungen vorgenommen werden, was die Erfassung von möglichst vielen Facetten der Kompetenzausprägung der Personen ermöglicht. Die empirische Konstruktion der Stufen erfolgt über die Entwicklung einer hinreichend großen Anzahl von Testaufgaben, welche zunächst pilotiert und in Hinblick auf die Aufgabenschwierigkeit skaliert werden. Anschließend werden die gemessenen Daten über die Inhalte der einzelnen Aufgaben der Skala zugeordnet. Eine detaillierte Beschreibung des Verfahrens der Aufgabenzuordnung zu den Kompetenzstufen kann dem Abschnitt 4.3.3 entnommen werden. Für eine ausführliche Betrachtung der empirischen Erstellung von Kompetenzstufenmodellen sei auf Hartig (2004) verwiesen.

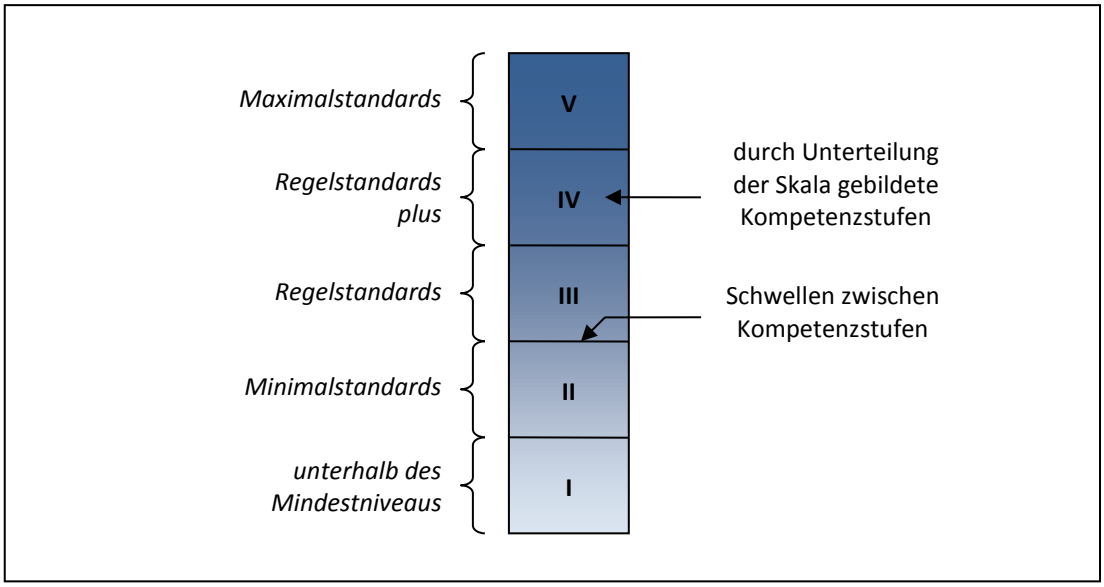

Abbildung 2: Kompetenzstufenmodell

Innerhalb eines Skalenabschnitts werden keine weiteren inhaltlichen Differenzierungen der Kompetenz vorgenommen, so dass die Schwellensetzung zwischen den einzelnen Stufen entscheidend für die Zuordnung einer Fähigkeit zu einem Kompetenzniveau ist (vgl. Hartig & Klieme, 2006, S. 134 f.). Die Modelle sind sachlogisch-graduell aufgebaut und zugleich entwicklungspsychologisch bedingt. Nach Ziener (vgl. 2006, S. 37 ff.) sind für die Konstruktion von sachlogisch-graduellen Stufen die Voraussetzungen oder Grundfertigkeiten für das Erreichen der nächsten Kompetenzstufe bedeutsam. Das Modell besteht aus diesem Grund aus folgerichtig aufeinander aufbauenden Kompetenzen, wobei beim kumulativen Kompetenzerwerb keine übersprungen werden darf. Die Berücksichtigung von Erkenntnissen der Entwicklungspsychologie bedeutet hingegen, dass beispielsweise bei Primarstufenschülern noch kein übertragendes oder symbolisches Denken vorausgesetzt werden kann. Daher ist jedes Niveau durch eine gewisse Qualität von Prozessen oder Handlungen gekennzeichnet (vgl. Klieme, et al., 2007, S. 22). Das Ziel besteht in dessen Konsequenz in der Zuordnung des Kompetenzerwerbsprozesses des Schülers zu den einzelnen Stufen und dies im Idealfall über seine gesamte Schullaufbahn hinweg. Die bereits konstruierten Kompetenzstufenmodelle sind jedoch noch nicht mit Längsschnittdaten empirisch geprüft, so dass abzuwarten bleibt, ob sich die Skalen für die Beschreibung von Lernentwicklungen als geeignet erweisen (vgl. Klieme, 2004, S. 3).

Die vom IQB erarbeiteten Kompetenzstufenmodelle orientieren sich zum einen an den bei PISA verwendeten Skalen, zum anderen wurde der Prozess der Skalierung mit den nationalen Begleitstudien von PISA und IGLU gekoppelt. Besonders zu bemerken ist hierbei die Unterteilung der Skala in fünf Stufenabschnitte, deren Erreichen wiederum mit den Bil-

dungsstandards der KMK verknüpft ist. Auf diese Weise wird eine Lösung für die Problematik der Regelstandards angestrebt (vgl. Abschnitt 2.2.2), indem eine Schülerleistung der Kompetenzstufe III so zu bewerten ist, dass sie dem *Regelstandard* entspricht und somit das durchschnittliche Lernziel erreicht ist (vgl. Abbildung 2). Aufgrund der Abstufung in zwei niedrigere Stufen können ebenso Mindeststandards generiert werden. Die Kompetenzstufe II entsprecht demnach dem *Mindeststandard*, währenddessen ein Schüler mit auf der Kompetenzstufe I verorteten Leistungen das Bildungsminimum verfehlt und den Mindeststandard nicht erreicht hat. Um den Lehrenden und Lernenden eine motivierende Richtung für die Weiterentwicklung aufzuzeigen, werden über die Regelstandards hinausgehend zwei weitere Kompetenzstufen definiert. Kompetenzen der Stufe IV werden als *Regelstandards plus* bezeichnet, während diejenigen der Stufe V als *Maximalstandards* zu charakterisieren sind und deren Erreichen die Erwartungen der Bildungsstandards übertrifft (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2009c, S. 6 ff.). Der Mittlere Schulabschluss, der Hauptschulabschluss sowie der Abschluss der Primarstufe streben jeweils das Erreichen der Regelstandards, das heißt die Kompetenzstufe III, an.

Für die erste Fremdsprache bildet der „Gemeinsame europäische Referenzrahmen für Sprachen: Lernen, lehren, beurteilen“ (GER) die Grundlage für das Kompetenzstufenmodell des IQB. Dieses Handlungsmodell ist eine Beschreibung der erforderlichen Leistung, „um eine Sprache für kommunikative Zwecke zu benutzen, und welche Kenntnisse und Fertigkeiten [die Lernenden] entwickeln müssen, um in der Lage zu sein, kommunikativ erfolgreich zu handeln“ (Europarat, 2001, S. 14). Indem dieser Kompetenzrahmen die Basis für den Spracherwerb in Europa bildet, sorgt er für Transparenz und Vergleichbarkeit der Leistungserwartungen. Die hier verwendeten Skalen sind ebenfalls sachlogisch-graduell aufgebaut und umfassen sprachliche und kognitive Kompetenzen. Die Skalen werden für die jeweiligen Kompetenzdimensionen in die drei Grundniveaus A, B und C unterteilt, welche wiederum zwei Teildimensionen umfassen (A1, A2, B1 etc). Das Niveau A entspricht der elementaren Sprachverwendung, das Niveau B der selbstständigen Sprachverwendung und das Niveau C kann als kompetente Sprachverwendung interpretiert werden (vgl. Köller, 2008a, S. 167 f.). Aufgrund der Ergebnisse der Skalierung durch das IQB mussten jedoch weitere Untergliederungen in A1.1, A1.2, A2.1 etc. vorgenommen werden. In dessen Konsequenz ergeben sich für die verschiedenen Abschlüsse unterschiedliche Zuordnungen zu den einzelnen Stufen. Während für den Mittleren Schulabschluss das Niveau B1.2 den Regelstandards entspricht, wird für den Hauptschulabschluss das Niveau A2.1 angestrebt (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2009d; 2009e). Ob diese vielfach unterteilten Skalen der einfacheren Kommunikation und Visualisierung der Kompetenzanforde-

rungen dienen, sei dahingestellt. Die Kritik von Keller & Ruf (vgl. 2005, S. 456), dass im GER ausschließlich fachliche Kompetenzen aufgegriffen worden seien und nirgends ein Bezug zu Selbst- und Sozialkompetenzen zu finden sei, lässt sich ebenfalls auf die Kompetenzstufenmodelle anderer Fachdisziplinen übertragen. Die Ursache für diese Problematik liegt in der bisher fehlenden Möglichkeit der empirischen Skalierung überfachlicher Kompetenzen, da sie sich nicht mit kognitiven Leistungen auf einer gemeinsamen Skala abbilden lassen.
Mittels der Skalierung im Zuge der Entwicklung von Kompetenzstufenmodellen konnte die Verteilung der Leistungen der betreffenden Schülerpopulation auf den jeweiligen Niveaustufen gemessen werden. Für die genauen Ergebnisse sei an dieser Stelle auf die Ausführungen in den einzelnen Veröffentlichungen der Kompetenzstufenmodelle verwiesen. Als ein besonders auffälliges Ergebnis ist anzuführen, dass 54,5 Prozent der Schüler der Klassenstufe 9, welche den Hauptschulabschluss anstreben, nicht die Mindeststandards in Mathematik erreichen, das heißt auf der Kompetenzstufe I zu verorten sind. Lediglich 19 Prozent dieser Schülergruppe wiesen Leistungen im Bereich der Regelstandards auf (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2009b, S. 39). Hier wird besonders die Notwendigkeit einer Optimierung der Lernprozesse deutlich. Die Bildungsstandards sollen der Verbesserung dieser Situation dienen, indem sie unter anderem als Grundbausteine für die Unterrichtsplanung herangezogen werden und sich ein kompetenzorientierter Unterricht in der schulischen Praxis etabliert.

Resümierend betrachtet dienen Kompetenzmodelle nicht allein der Testentwicklung in Form einer theoriegeleiteten Aufgabenentwicklung, sondern sollen zugleich als strukturierende Orientierungsinstrumente bei der konkreten Unterrichtsplanung und Diagnostik zum Einsatz kommen (vgl. Hartig, 2004, S. 74; Klieme & Leutner, 2006, S. 885). Sie tragen daher zu einer besseren Kommunikation und zu einer Transparenz der Leistungserwartungen bei. Als Unterstützungshilfe für die Leistungsdiagnostik kann die zu interpretierende Kombination von Kompetenzstrukturmodell und Kompetenzstufenmodell in Form eines Kompetenzrasters sinnvoll sein (vgl. Abbildung 3).
Für das Fach Mathematik werden beispielsweise im Strukturmodell sowohl sechs mathematische Kompetenzen als auch fünf mathematische Leitideen ausgewiesen. Hinzu kommen die fünf Niveaus des Kompetenzstufenmodells. Somit ergibt sich ein dreidimensionales Raster, in welchem die derzeitige Leistung sowie der Lernprozess insgesamt verortet werden kann. Im Beispiel der Abbildung 3 hätte der Schüler somit im Bereich der mathematischen Leitidee L4 und der mathematischen Kompetenz K3 die Kompetenzstufe IV erreicht, was den Regelstandards plus entspräche.

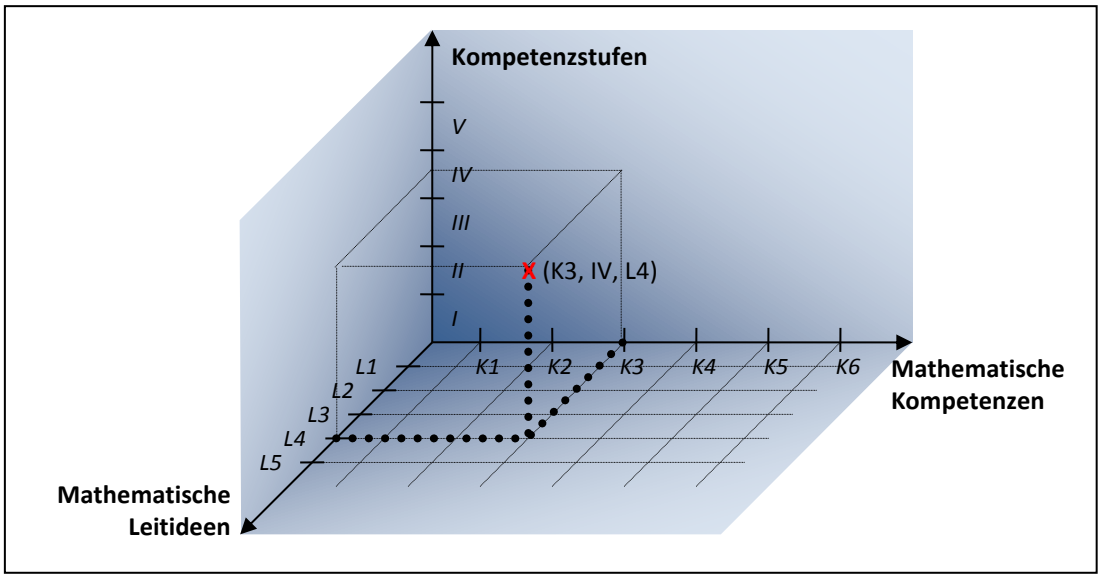

Abbildung 3: Kompetenzraster am Beispiel des Faches Mathematik

2.4 Kritik an den Bildungsstandards und den zugehörigen Kompetenzmodellen

Mit dem Beschluss zur Entwicklung und der darauf folgenden Formulierung der Bildungsstandards entbrannte zugleich eine intensive öffentliche Kritik an deren Konzeptionalisierung. Während sich einige Argumentationen mit der Implementierung sowie mit der Weiterentwicklung der Standards mittlerweile erübrigt haben, werden andere weiterhin in der aktuellen Fachliteratur vertreten und folgend in ihren Grundlinien knapp dargestellt.

Deutlich abzugrenzen seien die Bildungsstandards nach Heymann (vgl. 2005a, S. 24) vom Begriff der Allgemeinbildung. Während in den Standards nur fachliche Leistungen berücksichtigt werden, benötige der Mensch für eine umfassende Bildung ebenso sozialethische und personale Kompetenzen, so dass mit den Standards der Eindruck eines verkürzten Bildungsbegriffs entstehe. Criblez, et al. (vgl. 2009, S. 11) betonen in diesem Zusammenhang ebenfalls, dass eine Skalierung der zu erbringenden Leistungen dem pädagogischen Konzept der Individualisierung konträr entgegenstehe.

Zudem wurden nur die Kernfächer mit nationalen Standards versehen, so dass eine Vernachlässigung der anderen Disziplinen, die einen Großteil des zu unterrichtenden Fächerspektrums ausmachen, zu befürchten sei. Uhle (vgl. 2007, S. 47) verknüpft diese Konzentration auf wenige überprüfbare Fächer mit einer anstehenden Ökonomisierung der Bildung. Kritik an den teilweise unpräzisen Formulierungen der Standards, welche massive Interpretationsspielräume offen ließen und subjektive Interessenskonflikte der Konstruktionsteams offenbaren würden, äußert Drieschner (vgl. 2009, S. 60). „Die Unklarheit von Kompetenzbe-

schreibungen wird vor allem aus testpsychologischer Sicht zum Problem, insofern sie sich nicht in psychometrische Aufgaben überführen und über Vergleichsarbeiten messen lassen“ (ebd., S. 60). Hieraus leitet er die Forderung nach einer eindeutigen Operationalisierbarkeit der Standards ab (vgl. ebd., S. 61). Als weitere Anmerkung führen Criblez, et al. (vgl. 2009, S. 173) an, dass die Formulierung der Leistungserwartungen oftmals zu eng an den inhaltlichen Dimensionen der Lehrpläne angelehnt sei, so dass hier ein Mangel an Fokussierung zu erkennen sei.

Ein anderer Argumentationsstrang setzt an der Konzeption von abschlussbezogenen Standards an. Heid (vgl. 2006, S. 19 ff.) äußert sich dahingehend, dass mit den Standards eine überdimensionale Wirkungskette erwartet werde, welche sich als unrealistisch erweisen werde. Indem am Ende eines Lernabschnitts die erworbene Kompetenz in Form eines Leistungsergebnisses gemessen werde, könne hieraus noch keine Aussage über den bereits abgeschlossenen Lernprozess getätigt werden. Daher seien aufgrund eines erreichten Kompetenzstandes keine Schlussfolgerungen über die Qualität des Lehrens, die jeweilige Lehrerprofessionalität, die Qualität der Einzelschule oder gar des gesamten Bildungssystems möglich. Nach Regenbrecht (vgl. 2005, S. 17) müssten für eine solche Analyse und Bewertung zahlreiche weitere Indikatoren, wie die schulische Ausgangslage, das Schulprofil, das Schulklima, die Schulkultur, das soziale Lernen etc. berücksichtigt werden.

2.5 Kerncurricula

Im Zuge des Paradigmenwechsels in Richtung einer zunehmenden Outputsteuerung im Bildungswesen stellen die Bildungsstandards ein neuartiges Instrument für die Qualitätsentwicklung und -sicherung dar. Zugleich finden weitere Reformprozesse im Bereich der staatlichen Vorgaben statt, indem die herkömmlichen Lehrpläne allmählich von Kerncurricula abgelöst werden. Die Lehrpläne sind als die klassischen Instrumente einer Inputsteuerung anzusehen, indem sie festlegen, welche Lerninhalte in einem Fach einer bestimmten Jahrgangsstufe in einer spezifischen Schulform zu unterrichten sind (vgl. Klieme, et al., 2007, S. 44). Ausgerichtet sind sie in ihrer Konzeption an fachlichen, fachdidaktischen und pädagogischen Theorien. Somit bilden sie eine strukturierte Grundlage und eine Orientierung für die Unterrichtsplanung.

Problematisch erwies sich in diesem Zusammenhang zum einen die zu starke Detailgenauigkeit, die sich verbunden mit einer Überfrachtung der Lehrpläne einengend auf die pädagogische Freiheit und zugleich belastend für die Lehrpersonen auswirkte. Zum anderen repräsentieren die Lehrpläne nach Klieme, et al. (vgl. ebd., S. 91) stets eine vorgenommene

Selektion und Transformation von Bildungszielen in staatliche Vorgaben, so dass eine Normierung der Unterrichtsgestaltung und erwünschten Lernergebnisse intendiert gewesen sei. Mit den Lehrplänen sind jedoch noch keine Aussagen getroffen, wie diese Vorgaben tatsächlich genutzt werden. Sie erheben lediglich implizit die Forderung, dass alle Schüler die Ziele im gleichen Maße zu erreichen haben.

Die Einführung der Bildungsstandards fungierte als eine enorme Treibkraft für die Ablösung der Lehrpläne durch Kerncurricula. Wie der Begriff bereits aussagt, enthalten die Kerncurricula den „Kern“ eines Faches, das heißt das unentbehrliche Minimum der zu behandelnden Themenblöcke, Inhalte und Lernformen. Auch sie enthalten Beschreibungen der Lernziele in Form der Standards, spezifiziert nach dem Unterrichtsfach. Die Orientierungs- und Strukturierungsfunktion der herkömmlichen Lehrpläne wird auf die Kerncurricula übertragen. Diese sind jedoch durch eine enorme Reduzierung der Vorgaben gekennzeichnet. Auf diese Weise besteht die Möglichkeit einer individuellen fachlichen Vertiefung oder Erweiterung mittels einer selbstständigen, didaktisch begründeten Auswahl von Lerninhalten durch die Lehrperson. Curriculare Entscheidungen werden zunehmend in die Hände der einzelnen Lehrkörper bzw. in die Verantwortung der eigenständigen Schule übertragen, welche aufbauend auf dem staatlich vorgegebenen Kerncurriculum ein schulinternes Curriculum entwickeln kann (vgl. Vortmann, 2005, S. 126). Dies befördert wiederum eine thematische Profilierung der Schule. Hierbei sollte angemerkt werden, dass sich die Konzepte der Kerncurricula in den Bundesländern inhaltlich voneinander unterscheiden.

Die Kerncurricula sind nicht als Gegenstücke der Bildungsstandards zu interpretieren, sondern vielmehr als deren Ergänzung (vgl. Criblez, Oelkers, Reusser, Berner, Halbheer, & Huber, 2009, S. 117 f.). Gemeinsam bilden sie den Referenzrahmen für unterrichtliches Handeln, indem mithilfe der Kerncurricula eine weitere Präzisierung der Standards über die Benennung von praxisrelevanten Lerninhalten vorgenommen wird und die beiden Komponenten Wissen und Können des Kompetenzbegriffs gleichermaßen berücksichtigt werden (vgl. Böttcher, 2003, S. 155). Kerncurricula und Bildungsstandards sind aus diesem Grund als unverzichtbare Instrumente der Qualitätsentwicklung in einem Systemzusammenhang zu betrachten. Die Kerncurricula stellen das zentrale Gelenkstück des Inputs dar, währenddessen der Output als zu erwartende Lernergebnisse durch die Standards festgelegt ist. Zwischen diesen zwei Polen finden die Lernprozesse statt (vgl. Abbildung 4). Diese Prozessphase wird durch Unterstützungssysteme, wie Lehrerfortbildungen, ergänzt. Als viertes Instrument findet zum Abschluss des Lernprozesses die Überprüfung der Standards und der Schülerleistungen statt (vgl. Wolff, 2009, S. 3).

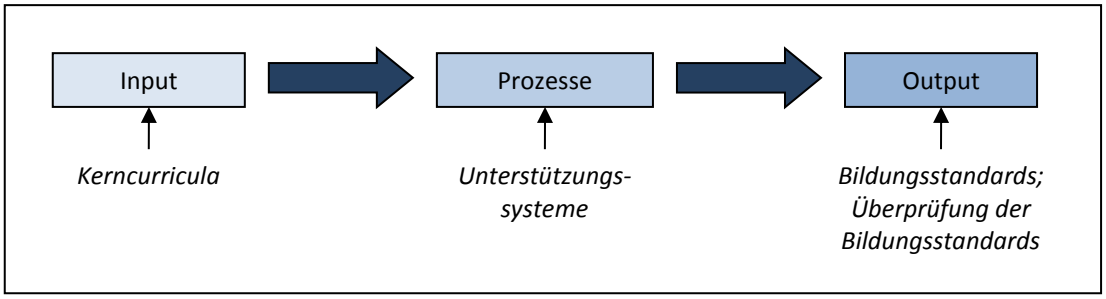

Abbildung 4:Zusammenhang zwischen dem Steuerungsmodell und den Instrumenten zur Qualitätssicherung- und Entwicklung

Für eine zu entfaltende Wirkung und eine Akzeptanz dieser Qualitätsinstrumente ist es zwingend erforderlich, dass die einzelnen Maßnahmen ein schlüssiges Gesamtkonzept mit einer übereinstimmenden Zielrichtung bilden. Steffens (vgl. 2009, S. 3) fasst diese Voraussetzungen mit den Termini Kohäsion, Konsistenz und Konkordanz zusammen. Bedeutsam sei demnach ein sukzessives, kumulatives Vorgehen bei der Implementierung von Kerncurricula und Bildungsstandards, welches durch eine inhaltliche Abstimmung und Gleichzeitigkeit gekennzeichnet sei.

2.6 Kompetenzorientierter Unterricht

Mit der Reduktion der staatlichen Vorgaben über die Implementierung von Bildungsstandards und Kerncurricula erhalten die Lehrkräfte wachsende Freiräume für ihre eigene Unterrichtspraxis. Hieraus ergibt sich die Chance, den Unterricht nicht primär an stofflichen Vorgaben, sondern zunehmend an den Lernausgangslagen und Erfordernissen der jeweiligen Schülergruppe zu orientieren (vgl. Höfer, Steffens, Diehl, Loleit, & Maier, 2009, S. 6). Angestrebt wird ein kumulativer Lernprozess, welcher nicht „träges“, sondern anwendbares Wissen und Können im Sinne des Erwerbs von Kompetenzen ermöglicht.

Der sogenannte „kompetenzorientierte Unterricht“ ist durch eine starke Schülerorientierung gekennzeichnet und richtet sich an der individuellen kognitiven Entwicklung der Lernenden aus (vgl. Klieme, et al., 2007, S. 50). Handlungsorientierung stellt hierbei ein zentrales Merkmal des Unterrichts dar. Zuvor erworbenes Wissen muss angewandt werden, um dieses Wissen in Können zu transformieren. Wie Klieme, et al. (vgl. ebd.,S. 78 f.) anmerken, zeichnet sich das Erreichen von nächsthöheren Kompetenzstufen durch eine zunehmende Prozeduralisierung des Wissens aus, indem die Schüler stets komplexere Anforderungssituationen zu bewältigen haben.

Beim kompetenzorientierten Unterricht gilt es, „Wissen und Können, fachliche, überfachliche und selbstregulative Kompetenzen gleichermaßen zu ‚vermitteln', um die Entwicklung kognitiver Strukturen auf Seiten der Schüler(innen) zu befördern, die zu kompetenten Operationen oder Handlungen befähigen" (Lersch, 2007, S. 439). Die Methodik des kompetenzorientierten Lernens besteht somit zum einen aus der Lehrmethode, bei welcher in spezifischen Lerngelegenheiten Wissen vermittelt wird, und zum anderen aus der Lernmethode, welche eine Anwendung des Wissens ermöglicht. Nur mittels dieser Kombination können fachliche und überfachliche Kompetenzen erworben werden (vgl. Lersch, 2010, S. 11). Dies geht mit veränderten Anforderungen an die Lehrerprofessionalität einher. Während sich die bisherige Unterrichtsplanung vorrangig an den Vorgaben des Lehrplanes orientierte, müssen nun ebenso die intendierten Wirkungen der Lehr-Lern-Prozesse berücksichtigt werden (vgl. Steffens, 2009, S. 2). Demzufolge muss der kompetenzorientierte Unterricht vom angestrebten Endergebnis ausgehend langfristig und kontinuierlich geplant werden (vgl. Kultusministerkonferenz, 2010, S. 4).

Für einen kumulativen Kompetenzerwerb sind spezifische Lernarrangements erforderlich, welche an den Lernausgangslagen der Schüler ansetzen, sich für die Aneignung der jeweiligen Kompetenz in besonderer Weise eignen sowie ein nachhaltiges und anschlussfähiges Lernen ermöglichen (vgl. Höfer, Steffens, Diehl, Loleit, & Maier, 2009, S. 6). Des Weiteren ist es unabdingbar, dass die einzelnen Lerngelegenheiten an die jeweiligen Kompetenzstufen angepasst sind, so dass das tatsächliche Lernpotenzial der Schüler ausgeschöpft werden kann (vgl. Drieschner, 2009, S. 80). Somit kommt der Lehrperson größere Freiheit in der unterrichtlichen Planung zu; andererseits wird sie mit der gewachsenen Verantwortung konfrontiert, mit Hilfe von spezifischen Lernangeboten die Schüler individuell und nachhaltig zu fördern (vgl. Helmke U. , 2005, S. 449).

Hiermit seien lediglich die Grundgedanken des kompetenzorientierten Unterrichts dargestellt. Für spezifische didaktische Ansätze sei auf Lersch (2006; 2007; 2010) Lange (2005) und Ziener (2006) verwiesen.

3 Standardisierte Kompetenztests als Instrumente zur Überprüfung der Bildungsstandards

Die Bildungsstandards geben in Form erwünschter Ergebnisse der schulischen Lehr-Lern-Prozesse den zu erreichenden Output vor. Dies impliziert die Notwendigkeit, an spezifischen Schnittstellen der Schullaufbahn zu überprüfen, inwiefern der geforderte Output tatsächlich vorliegt. Konkret bedeutet dies eine Testung der vorhandenen Schülerleistungen, gemessen an den Vorgaben der Bildungsstandards. Bereits in der Expertise empfahlen Klieme, et al. (vgl. 2007, S. 23) die kontinuierliche Überprüfung der Bildungsstandards mithilfe von Aufgabenstellungen und Verfahren, mit denen das Kompetenzniveau, das die Schüler erreicht haben, empirisch zuverlässig erfasst werden kann.
Als einen zentralen Anstoß für objektivierte Outputmessungen, zu denen unter anderem auch die zentralen Abschlussarbeiten zählen, betont Schirp (vgl. 2006b, S. 424) die neue Steuerungsphilosophie und sieht damit wiederum die bereits bekannten Begründungslinien verbunden:

- Ökonomisierung der Bildungssysteme, infolge derer die Schulen daran gemessen werden, was sie ausgedrückt in den Leistungsergebnissen ihrer Schüler tatsächlich leisten,
- Notwendigkeit eines objektiven Leistungsvergleichs, durch den die Stärken und Schwächen der Bildungssysteme erkannt und daraufhin Umsteuerungen vorgenommen werden können,
- Leitbild einer selbstständigen Schule, die sich anhand ihres Abschneidens in Leistungstests profilieren und Rechenschaft ablegen kann.

Hieran ist erneut der Systemzusammenhang der einzelnen Reformentwicklungen erkennbar, welche sich teilweise unabhängig voneinander entwickelten haben und eine gebündelte Wirkkraft entfalten. In diesem Kontext stellen auch Dedering, et al. (vgl. 2007, S. 408 ff.) in einer Untersuchung zu zentralen Leistungsmessungen fest, dass bereits vor PISA und der Formulierung von Bildungsstandards in den einzelnen Bundesländern Leistungsvergleichsuntersuchungen initiiert worden waren. Sie gelangen zu dem Fazit, dass PISA nicht das auslösende Moment gewesen sei, sondern vielmehr als Begründung bzw. als Katalysator für die Ausweitung von standardisierten Lernstandsmessungen verschiedenster Formen fungiert habe.
Als Beispiel für eine solche Entwicklung kann das Bundesland Thüringen herangezogen werden: Zur Sicherung der Bildungsqualität sind zunächst zentrale Abschlussarbeiten für jeden Schulzweig etabliert worden. Als Konsequenz aus PISA wurde dieses Konzept mit

Vergleichstests in den Klassenstufen 3 und 6 sowie durch die Besondere Leistungsfeststellung in der Klassenstufe 10 des gymnasialen Bildungsgangs erweitert. Mit der Einführung der Bildungsstandards wurden die Vergleichstests wiederum zu Kompetenztests transformiert und die achte Klassenstufe ebenfalls in dieses Konzept integriert (vgl. Korngiebel, 2009, S. 34).

Um zu gewährleisten, dass solche Messmodelle tatsächlich ihrer Funktion als Überprüfungsinstrument der Bildungsstandards nachkommen, müssen sie klaren Anforderungen genügen. Dabei ist vor allem die Abstimmung auf die empirisch gesicherten Kompetenzmodelle notwendig, um darauf aufbauend die Lernergebnisse differenziert erfassen zu können. Es ist aus diesem Grund legitim, solche Testmodelle zunächst generalisierend als Kompetenztests zu charakterisieren, wie im Folgenden näher ausgeführt wird.

3.1 Präzisierung des Terminus: „Kompetenztest“

Der Begriff „Kompetenz“ wurde bereits in Abschnitt 2.3.1 explizit erläutert. Laut jener Definition bezeichnen Kompetenzen erlernbare Fähigkeiten und Fertigkeiten, die sowohl kognitiver als auch motivationaler, volitionaler und sozialer Natur sein können.

Ein Test hingegen wird nach Lienert folgendermaßen definiert: „Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbaren Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung“ (1969, S. 7). Ein Test ist jeweils von den zwei Phasen der Lösungspraxis und der Auswertung determiniert. In der ersten Phase werden die schriftlich oder mündlich dargelegten Lösungen des Probanden erfasst. Gedachte Lösungen können in Tests nicht ermittelt werden. Anschließend erfolgt die Kodierung und Auswertung durch standardisierte Verfahren. Hierbei wird eine Reduktion der Messdaten vorgenommen, indem Lösungsteile, die nicht mit dem vorgegebenen Kodierungsmuster übereinstimmen, bei der Auswertung keine Beachtung finden (vgl. Meyerhöfer, 2005, S. 20).

Auf die empirische Bildungsforschung bezogen ist das Ziel eines Kompetenztests die valide, genaue und objektive Reflexion der Kompetenzausprägung von Schülern. Dies ist zwar einerseits abzugrenzen vom Begriff der Evaluation, der eine Bewertung von Maßnahmen bezüglich zuvor festgelegter Ziele beinhaltet (vgl. Klieme & Leutner, 2006, S. 881), andererseits können die Tests ein Instrument innerhalb einer Evaluation darstellen. Kompetenztests in der Schule erheben demzufolge auf möglichst objektive Art und Weise, ob die Schüler ein bestimmtes Unterrichtsziel erreicht haben und ob hieraus schlussfolgernd die Lern-

prozesse als erfolgreich betrachtet werden können (vgl. Boes, 2003, S. 94). Diese Tests können als psychometrische Leistungsmessungen charakterisiert werden, die sich von subjektiven Persönlichkeitsmessungen abgrenzen. Zudem vermögen Kompetenztests bislang nicht die potentielle Maximalleistung eines Schülers aufzuzeigen, da nicht alle denkbaren Fähigkeiten in ihren Anforderungsstufen berücksichtigt werden können. Statt der Leistungsgrenze wird somit die derzeitige Leistungsfähigkeit eines Probanden gemessen (vgl. Tresch, 2007, S. 42 f.).

Kompetenztests können entweder als externe standardisierte Instrumente konstruiert sein oder selbst entwickelte Beurteilungskonzepte der Lehrperson darstellen, die beispielsweise wie die traditionellen Klassenarbeiten im Klassenverband durchgeführt werden. In Hinblick auf die Überprüfung von Bildungsstandards wird ersteres Verfahren im Zentrum stehen, bei dem eine klare und nachprüfbare Zuordnung zu Kompetenzen vorgenommen werden kann. Dies beinhaltet einerseits die Beschreibung erreichter Teilkompetenzen und ihrer Verknüpfung zu den jeweiligen Niveaustufen, andererseits sollte die Qualität der Lernprozesse einbezogen werden. Es reicht für die Lehrkraft nicht aus zu wissen, ob ein Schüler eine spezifische Kompetenz erworben hat, sondern es ist ebenso bedeutsam zu erkennen, auf welchem Weg er sie erlangt hat und welche Streuung sein Ergebnis zu Vergleichswerten aufweist. Der Unterschied zwischen Kompetenztests und bisherigen Lernstandsmessungen besteht darin, dass nicht kurzfristig erlerntes Wissen, sondern erworbene Fähig- und Fertigkeiten abgefragt werden, die durch ein nachhaltiges Lernen gekennzeichnet sind (vgl. Lersch, 2006, S. 32). Zusätzlich zeichnet sich ein qualitativ brauchbares Messinstrument dadurch aus, dass eine klare Interpretation des Messergebnisses erfolgt und über Rückmeldungen konkrete Handlungsmöglichkeiten ausgewiesen werden (vgl. Klieme & Leutner, 2006, S. 877 ff., 886).

3.2 Formen standardisierter Kompetenztests

Derzeit existiert in der Bundesrepublik eine Vielzahl an verschiedenen Testkonzepten zur Kompetenzmessung, welche bereits regelmäßig durchgeführt werden oder sich noch im Prozess ihrer Entwicklung und wissenschaftlichen Fundierung befinden. Der Einsatz eines Testmodells sollte stets an den Zweck und an das Ziel der Messung angepasst sein. Ebenso sollte das Untersuchungsdesign an den Informationsbedürfnissen der jeweiligen Adressaten ausgerichtet werden. Nur auf Basis dieser Bedingungen können die Ergebnisse später sinnvoll genutzt werden und produktiv Verwendung finden. Hieraus ergeben sich generelle

Kriterien, anhand derer eine Abgrenzung der verschiedenen Testmodelle voneinander vorgenommen werden kann:

- *Summative vs. formative Messung*

  Die summative Leistungsmessung beinhaltet eine bilanzierende Ermittlung des erreichten Lernstandes zu einem vorab festgelegten Zeitpunkt anhand spezifischer Kriterien (vgl. Criblez, Oelkers, Reusser, Berner, Halbheer, & Huber, 2009, S. 107). Hiermit ist oftmals das Ziel einer vergleichenden Darstellung der Leistungen verbunden. Indem der Blick auf das Resultat eines Lernabschnittes gerichtet ist, werden implizit die Wirkungen vorangegangener Lernprozesse bewertet, so dass der summativen Messung ein retrospektiver Blickwinkel innewohnt. Der formative Kompetenztest fokussiert hingegen stärker den eigentlichen Lernprozess und ermittelt individuelle Lernfortschritte und Zwischenergebnisse. Damit ist ein förderdiagnostischer Ansatz verknüpft, bei dem das Ziel der Messung in der Optimierung des weiteren Lernprozesses besteht. Formative Tests haben somit einen entwickelnden prospektiven Charakter (vgl. Maier, 2010b, S. 48; Tresch, 2007, S. 58 f.)

- *Querschnitt vs. Längsschnitt*

  Ein weiteres Kriterium ist das Untersuchungsdesign. Diesbezüglich gibt es Querschnitt- und Längsschnittuntersuchungen. Erstere erfasst einmalig die Leistung eines Schülers in spezifischen Kompetenzbereichen und eignet sich daher besonders bei einer großen Probandenzahl und bei der Betrachtung bestimmter Jahrgänge. Im Falle einer rhythmischen Wiederholung der Tests wird somit jeweils eine neue Schülergruppe erfasst. Eine Querschnittsmessung ist oftmals zugleich eine summative Messung. Bei einer längszeitlichen Betrachtung steht hingegen die Leistungsentwicklung eines Schülers oder einer Lerngruppe über mehrere Messzeitpunkte hinweg im Blickfeld, so dass diese Testmethode stark mit formativen Ansätzen korreliert (vgl. Tresch, 2007, S. 50).

- *Reichweite des Tests*

  Die Messungen sind zudem determiniert durch den Kreis der zu erreichenden Probanden. Die Kompetenztests können von einzelnen Schülern oder den Klassenverbänden bis hin zu großflächig teilnehmenden internationalen Schülerpopulationen reichen.

- *Vollerhebung vs. Stichprobe*

  Bei der Messung von Lernständen kann einerseits eine Vollerhebung durchgeführt werden, indem beispielsweise die Leistungen aller Schüler eines Jahrgangs getestet werden. Eine andere Möglichkeit besteht in der Auswahl einer genügend großen

Stichprobenzahl, welche wiederum die Gesamtheit ihrer Schülerpopulation repräsentiert (vgl. ebd., S. 50).

- *Testbereiche*
  Bezüglich des zu testenden Gegenstandes muss zuvor festgelegt werden, welche spezifischen Kompetenzen überprüft werden.

Die Kompetenztests können verschiedenen Funktionen, wie der Förderung, Selektion, Benotung oder dem Monitoring dienen (vgl. Hartig & Jude, 2007, S. 17). Unter Berücksichtigung der soeben genannten Kriterien werden im Folgenden die möglichen Testinstrumente hinsichtlich ihrer jeweiligen Funktion voneinander differenziert erläutert. Es kann zwischen *Bildungsmonitoring*, *Schulmonitoring und -entwicklung* sowie *Individualdiagnostik* unterschieden werden.

3.2.1 Tests zum Bildungsmonitoring

Großflächige standardisierte Leistungsmessungen, auch Large-Scale-Assessments genannt, dienen vorrangig dem Bildungsmonitoring. Über eine Bestandsaufnahme der erreichten Kompetenzen in Form einer Querschnittsmessung stellen sie den Output der Lernprozesse vergleichend zu einer größeren Zahl weiterer Schülerpopulationen dar. Die individuelle Schülerleistung wird hierbei nicht betrachtet. Vielmehr repräsentieren die Messdaten die gesamte Schülerschaft. Aus diesem Grund werden auch Lernprozesse nicht berücksichtigt. Die Bestandsaufnahme durch Large-Scale-Assessments ermöglicht es, fundiertes Wissen über Qualität, Defizite und Entwicklungsmöglichkeiten aufzuzeigen (vgl. Husfeldt, 2004, S. 500 ff.). Weiterführend kann dieses Steuerungswissen Hinweise für konkrete Handlungsoptionen im Rahmen der Qualitätsentwicklung des Bildungssystems liefern (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 9). Dedering, et al. (vgl. 2007, S. 409) räumen in diesem Zusammenhang allerdings ein, dass bislang ungeklärt ist, ob und inwiefern die derzeitigen Large-Scale-Assessments diesem Anspruch gerecht werden. Gleichzeitig erfüllt diese Form von standardisierten Kompetenztests eine Rechenschaftsfunktion, indem sie einerseits Aussagen bezüglich der Erfüllung des gesetzten Bildungsauftrages zulassen, andererseits Informationen über die Effektivität des Bildungssystems sowie über die Wirkkraft initiierter Reformmaßnahmen generieren (vgl. Husfeldt, 2004, S. 501; Klieme, et al., 2007, S. 102). Diese summativen Leistungstests sind somit eindeutig an die Bildungsadministration und -politik adressiert.

Die Large-Scale-Assessments werden hinsichtlich ihrer Reichweite unterschieden. Zu den internationalen Testmodellen, welche in Deutschland durchgeführt werden, zählen TIMSS, PISA und PIRLS (Progress in International Reading Literacy Study)/ IGLU. Hieran gekoppelt werden in einem zeitlichen Rhythmus von fünf bis sechs Jahren nationale Erhebungen in Form eines Ländervergleichs durchgeführt. Diese stichprobenartigen Messungen werden vom IQB entwickelt und sind eng an den Kompetenzen ausgerichtet, so dass sie primär der Überprüfung von Bildungsstandards in Form eines Bildungsmonitorings dienen (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2007, S. 12; Kultusministerkonferenz, 2006). Hinzu kommen diverse regionale Testdesigns, wie DESI (Deutsch-Englisch-Schülerleistungen-International) zur Messung der sprachlichen Leistungen in den Fächern Deutsch und Englisch. Large-Scale-Assessments können sowohl fachlich kognitive als auch überfachliche Kompetenzen (zum Beispiel Problemlösekompetenz bei PISA) erfassen und werden in den Kernfächern Deutsch, Mathematik und der ersten Fremdsprache sowie in den Naturwissenschaften durchgeführt.

3.2.2 Tests zur Individualdiagnostik

In Hinblick auf die individuelle Förderung eines Schülers existieren verschiedenste Testkonzepte, welche möglichst umfassend und tiefgründig den Fähigkeitsstand des Lernenden ermitteln sollen. Es werden hierbei wissenschaftlich erprobte Tests eingesetzt, welche je nach dem zu untersuchendem Testgegenstand sowohl fachliche als auch überfachliche Kompetenzen berücksichtigen können und anschließend anhand festgelegter Kriterien extern ausgewertet werden. Dieser Typus fungiert daher als eine Grundlage für weitere pädagogische Entscheidungen, wie der Erstellung von individuellen Förderplänen oder der Vorbereitung von Schullaufbahnentscheidungen (vgl. Leutner, Fleischer, Spoden, & Wirth, 2007, S. 151). Das Testdesign ist meist formativ und mit Längsschnittelementen verbunden. Die Lehrperson ist der Hauptadressat, da sie spezifische Daten über den Kompetenzstand ihrer eigenen Schüler gewinnt und somit Konsequenzen für ihr weiteres unterrichtliches Handeln ableiten kann. Ein bedeutsamer Bestandteil der Individualdiagnostik ist aus diesem Grund die Qualität des Informationsgehalts der Ergebnisrückmeldung. Als Beispiel für einen individualdiagnostischen Kompetenztest kann BeLesen genannt werden.

Weitere Testkonzepte stellen die zentralen Abschlussarbeiten und die Parallelarbeiten dar. Die zentralen Abschlussarbeiten sind extern entwickelte Leistungsüberprüfungen, welche alle Schüler eines Bundeslandes zum Abschluss ihrer Schullaufbahn zum gleichen Zeitpunkt absolvieren müssen (vgl. Hovestadt & Keßler, 2005, S. 9). Die Bildungsstandards für die

Sekundarstufen I und II werden zukünftig als Basis für diese Tests dienen. Obwohl die Leistungen eines jeden Schülers individuell ausgewertet werden, muss eine Einordnung in die Kategorie „Individualdiagnostik“ an dieser Stelle mit großem Vorbehalt erfolgen. Nicht die weitere Lernentwicklung soll mit den zentralen Abschlussarbeiten optimiert werden, sondern es erfolgt vielmehr eine Zertifizierung und sogleich eine Selektion (vgl. van Ackeren & Bellenberg, 2004, S. 134 f.). Die Zentralisierung der Abschlussarbeiten soll vorrangig eine gerechte Bewertungsgrundlage sowie eine bundeslandweite Vergleichbarkeit der Schulabschlüsse bewirken. Die zentralen Abschlussarbeiten beinhalten allerdings keine empirisch erprobten Testaufgaben und werden nicht wissenschaftlich ausgewertet. Sie sollten zwar an den Kompetenzen erstellt und individuell ausgewertet werden, aber einer Charakterisierung als Kompetenztest im Sinne des Abschnitts 3.1 halten sie nicht stand.

Ähnlich lassen sich die Parallelarbeiten beschreiben, welche in einigen Bundesländern in den Schulen verpflichtend eingeführt wurden. Diese von den Lehrkräften selbst entwickelten Tests werden in allen Parallelklassen eines Jahrgangs durchgeführt und anschließend intern ausgewertet. Ein schulübergreifender Austausch findet hierbei meist nicht statt. Wie bei den traditionellen Klassenarbeiten kann auf diese Weise eine Individualdiagnostik der Schüler vorgenommen werden. Zusätzlich wird mit den Parallelarbeiten die Hoffnung verknüpft, den innerschulischen fachlichen Diskurs über curriculare Inhalte, Unterrichtsmethoden und Bewertungsgrundlagen zu intensivieren. Insofern sind sie ebenso im Bereich der Schulentwicklung zu verorten. Indem sie oftmals mit einer Benotung der dargelegten Leistung für den Schüler verbunden sind, kommt an dieser Stelle wiederum der Aspekt der Selektion zum Tragen. Zu betonen ist beim Typus der Parallelarbeiten, dass es sich um intern konzipierte Tests handelt. Daher kann an dieser Stelle nicht automatisch angenommen werden, dass die Arbeiten an den Bildungsstandards ausgerichtet sind und sie tatsächlich Kompetenzen überprüfen. Die Bezeichnung der Parallelarbeiten als Kompetenztests ist somit an dieser Stelle eher zu unterlassen.

3.2.3 Tests zur Schul- und Unterrichtsentwicklung

Um einzelnen Schulen einen Vergleich der Fachleistungen ihrer Schüler mit sozial ähnlichen Bezugssystemen zu ermöglichen, werden wissenschaftlich entwickelte Vergleichsarbeiten durchgeführt, die nachfolgend im Zentrum dieser Arbeit stehen. Mittels Rückmeldungen der ermittelten Ergebnisse auf Klassenebene dient dieser Testtypus vorrangig der Bestandsaufnahme sowie der Weiterentwicklung von Schule und Unterricht. Die Lehrkraft sieht sich als Hauptadressat mit der Aufgabe konfrontiert, aus den Ergebnissen Konsequen-

zen für ihr unterrichtliches Handeln zu ziehen. Zugleich ordnet die KMK die Vergleichsarbeiten als ein Instrument des Bildungsmonitorings ein (vgl. Kultusministerkonferenz, 2006). Die Vergleichsarbeiten stellen somit innerhalb der Kompetenztests eine Art Mischform zwischen den beiden Polen Bildungsmonitoring und Individualdiagnostik dar, indem sie einerseits Steuerungswissen generieren und andererseits individuelle Förderungsmaßnahmen auslösen sollen. Die Charakterisierung als eine Mischform bestärkt sich auch bei Betrachtung der weiteren Unterscheidungskriterien, da die Tests sowohl summative als auch formative Elemente enthalten: Zum einen sollen mit den Vergleichsarbeiten Qualitätsprozesse angestoßen werden; zum anderen ist die Auswertung produktorientiert und lässt die Lernprozesse unberücksichtigt (vgl. Maier, 2010b, S. 50). Weil die Vergleichsarbeiten zugleich Monitoring-Komponenten aufweisen, tritt die Bildungsverwaltung neben den Lehrkräften als ein zusätzlicher Adressat auf. Bezüglich des Testdesigns können klare Aussagen getroffen werden: Die Vergleichsarbeiten werden flächendeckend in den einzelnen Bundesländern als Querschnittstest durchgeführt.

Diese Beschreibung der Vergleichsarbeiten kann lediglich als eine Kurzcharakterisierung für die Abgrenzung zu anderen Konzepten der Kompetenzmessung interpretiert werden. Bevor die Vergleichsarbeiten hinsichtlich ihrer Implementierung und ihren Funktionen vertieft erläutert werden (vgl. Abschnitt 4), erfolgt eine zusammenfassende Übersicht der vorgestellten Formen von Kompetenztests mit der Intention, die Abgrenzungen bzw. Überschneidungen der jeweiligen Testkonzepte voneinander grafisch zu verdeutlichen (vgl. Tabelle 1).

Kriterien	Formen standardisierter Kompetenztests				
	Tests zum Bildungsmonitoring			Tests zur Schul- und Unterrichtsentwicklung	Tests zur Individualdi-agnostik
Reichweite	interna-tional	natio-nal	regional	regional	regional
Beispiel	PISA	PISA-E	DESI	**Vergleichsarbeiten**	BeLesen
formativ vs. summativ	summativ			summativ sowie formativ	formativ
Querschnitt vs. Längs-schnitt	Querschnitt			Querschnitt	Längsschnitt
Vollerhebung vs. Stichpro-be	repräsentative Stichprobe			meist Vollerhebung (in einigen Bundesländern freiwillige Teilnahme)	nur geringe Anzahl an Testprobanden
Testbereiche	fachliche und überfachliche Kompetenzen			fachliche Kompetenzen	fachliche und über-fachliche Kompeten-zen
Adressat	Bildungsadministration und -politik			Schulleitung, Lehrkräfte, Bildungsadministration	Lehrkräfte

Tabelle 1: Differenzierung der verschiedenen Formen standardisierter Kompetenztests

4 Vergleichsarbeiten

4.1 Implementierung von Vergleichsarbeiten

Wie bereits dargelegt wurde, begann die Einführung standardisierter Leistungsmessungen bereits vor der Entwicklung der Bildungsstandards. Dies kann ebenso für die Vergleichsarbeiten konstatiert werden, welche in Folge von PISA und IGLU seit 2002 starken Eingang in das Bildungssystem fanden. Mit dem Beschluss zur Festlegung nationaler Bildungsstandards erhielten die Vergleichsarbeiten als ein mögliches Instrument zur Überprüfung der gestellten Leistungsanforderungen neues Gewicht und wurden massiv ausgeweitet. Die KMK fasste bezüglich dieser Form externer Leistungsmessung bereits am 17./18. Oktober 2002 in Würzburg den folgenden Beschluss: „[D]ie Länder [werden] in landesweiten bzw. länderübergreifenden Orientierungs- oder Vergleichsarbeiten überprüfen, in welchem Umfang die vereinbarten Standards tatsächlich erreicht werden. Ziel dieses Verfahrens soll es sein, eine Qualitätssicherung zu gewährleisten, sich darüber länderübergreifend auszutauschen und es den Schülerinnen und Schülern in allen Ländern der Bundesrepublik Deutschland zu ermöglichen, in allen Bildungsgängen über individuelle Förderung die gesetzten Ziele zu erreichen“ (Kultusministerkonferenz, 2002b). Aus diesem Beschluss geht bereits hervor, dass die Vergleichsarbeiten mit den beiden Polen *Bildungsmonitoring* sowie *Individualdiagnostik* assoziiert wurden. Im Rahmen der KMK-Gesamtkonzeption für Maßnahmen zur Feststellung der Leistungsfähigkeit des Bildungssystems und der Schulen wurden 2006 die Vergleichsarbeiten erneut als ein Instrument des Bildungsmonitorings vorgestellt (vgl. Kultusministerkonferenz, 2006).

Das Testkonzept der Vergleichsarbeiten findet seinen Ursprung im Projekt „VERgleichsArbeiten in der Grundschule“ (VERA), welches 2002 in Kooperation der Universität Koblenz/ Landau mit dem rheinland-pfälzischen Ministerium für Bildung, Wissenschaft, Jugend und Kultur initiiert wurde. 2004 begannen die ersten Testdurchläufe in der vierten Jahrgangsstufe in sieben Bundesländern. Mit der Einführung der nationalen Bildungsstandards entstanden einzelne Testkonzeptionen für weitere Klassenstufen, welche als bundeslandinterne bzw. bereits ansatzweise als bundeslandübergreifende Maßnahmen durchgeführt wurden. Im Sinne von Vereinheitlichung und Transparenz wurde anschließend das IQB mit der wissenschaftlichen Testentwicklung beauftragt. Die Organisation und Auswertung der Tests finden weiterhin in Eigenverantwortung der Landesinstitute statt (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2010).

Die vom IQB entwickelten Vergleichsarbeiten werden in den Klassenstufen 3 und 8 durchgeführt (kurz: VERA 3 und VERA 8) und ersetzen somit die länderspezifischen Vorgänger-

modelle. Die Tests knüpfen hierbei eng an die Bildungsstandards und die Kompetenzmodelle an. Aus diesem Grund werden auch lediglich Fächer überprüft, in welchen nationale Bildungsstandards vorliegen. Dies bedeutet für die Primarstufe eine Leistungsmessung in Deutsch und Mathematik, währenddessen bei VERA 8 die erste Fremdsprache hinzukommt. Da die Bildungsstandards abschlussbezogen konstruiert worden sind, erfolgte die Konzeption der Vergleichsarbeiten für die unmittelbar vorangehenden Klassenstufen. Auf diese Weise soll eventueller Entwicklungsbedarf rechtzeitig aufgedeckt und Fördermaßnahmen eingeleitet werden können, so dass die geforderten Standards bei Abschluss der Schullaufbahn tatsächlich erreicht werden. Die Vergleichsarbeiten stellen in diesem Sinne eine gewisse Zwischenkontrolle der Schülerleistungen in Hinblick auf mittelfristig zu erbringende Anforderungen dar.

An der Durchführung von VERA 3 und VERA 8 beteiligen sich derzeit alle 16 Bundesländer. Seit 2010 setzen auch Schulen der deutschsprachigen Gemeinschaft in Belgien und Südtirol Bestandteile von VERA 3 ein. Zudem existiert in einigen Bundesländern eine weitere Vergleichsarbeit für die Klassenstufe 6, hier zur Vereinfachung VERA 6 genannt. VERA 6 wird jedoch nicht vom IQB entwickelt, sondern in einer interdisziplinären, länderübergreifenden Entwicklergruppe der sechs Teilnehmerländer Hamburg, Hessen, Mecklenburg-Vorpommern, Sachsen, Schleswig-Holstein und Thüringen. Die Durchführung der Vergleichsarbeiten erfolgt in den Bundesländern mehrheitlich flächendeckend und verpflichtend. In einigen Bundesländern beruht die Implementierung bislang auf Freiwilligkeit, um die Akzeptanz der Tests in den Schulen zu erhöhen. Langfristig wird jedoch vermutlich auch hier eine Verpflichtung zur Lernstandsmessung vorgenommen werden.

Der länderspezifischen Vorgängermodelle zu VERA 3 und VERA 8 sowie der föderalistischen Autonomie der Bundesländer hinsichtlich der Durchführung und Auswertung der Vergleichsarbeiten ist es geschuldet, dass die Bezeichnungen der Tests zwischen den einzelnen Ländern enorm variieren. VERA 8 wurde beispielsweise in Berlin als Vergleichsarbeit, in Hessen als Lernstandserhebung und in Thüringen als Kompetenztest implementiert, obwohl es sich in allen Bundesländern um einen identischen Test handelt (vgl. Korngiebel, 2009, S. 34). Für die weiteren Ausführungen dieser Arbeit wird zur Vereinfachung der Terminus „Vergleichsarbeiten“ weiterhin verwendet. Begründet wird dies damit, dass auch das IQB als testentwickelndes Institut das Testkonzept als „Vergleichsarbeiten“ tituliert. Daher fungiert die Bezeichnung „Vergleichsarbeiten“ in dieser Arbeit als ein Oberbegriff, welcher die länderspezifischen Betitelungen von VERA 3 und VERA 8 inkludiert. Im Sinne einer Vereinheitlichung wird VERA 6, welche mit identischen Zielsetzungen wie VERA 3 und VERA 8 be-

haftet ist, in diesem Rahmen ebenfalls hinzugezählt, obwohl die Testentwicklung nicht durch das IQB erfolgt.

4.2 Ziele und Funktionen von Vergleichsarbeiten

Die Doppelfunktionalität von Vergleichsarbeiten zwischen den Spannungsfeldern *Bildungsmonitoring* und *Individualdiagnostik* wurde bereits in dem Abschnitt 3.2.3 angedeutet. Zur Präzisierung dieser Problematik werden in den folgenden Ausführungen die verschiedenen Funktionen systematisiert, welche von bildungspolitischer sowie von wissenschaftlicher Seite den Vergleichsarbeiten zugeschrieben werden. Eine Übersicht dieser Funktionen ist in Tabelle 2 zu ersehen. Es ist zu untersuchen, inwiefern diese vielfältige Funktionenzuschreibung zu den Vergleichsarbeiten einen Widerspruch oder eine sinnvolle Integration von Qualitätsaspekten darstellt.

Funktionen von Vergleichsarbeiten			
Individual-diagnostik	**Schul-und Unterrichtsentwicklung**		**Bildungsmonitoring**
	Qualitätssicherung (summativ)	Qualitätsentwicklung (formativ)	
• Diagnostizieren und Fördern auf Schülerebene ?	• Schulevaluation • Bestandsaufnahme • Verortung • Rechenschaftslegung	• Innovationsanreiz für einen kompetenzorientierten Unterricht • Diagnostizieren und Fördern (auf Klassenebene) • Weiterentwicklung der Lehrerprofessionalität (Selbstreflexion, diagnostische Kompetenz, Orientierung an Bildungsstandards, Medienkompetenz) • Intensivierung der Fachgruppen- und Fachkonferenzarbeit	• Bestandsaufnahme • Feststellung von schulischem Unterstützungsbedarf • Unterstützung der Implementierung von Bildungsstandards • Transparenz der in den Bildungsstandards formulierten Anforderungen • Validierung der Kompetenzmodelle

Tabelle 2: Funktionen von Vergleichsarbeiten

Die EMSE, welche in Form einer Kooperation aller für die Durchführung von Vergleichsarbeiten zuständigen Landesinstitute zusammentritt, führt als zentrale Funktion dieser Form der Leistungsmessung die *Anregung von Schul- und Unterrichtsentwicklung* an (vgl. Netzwerk Empiriegestützte Schulentwicklung, 2006, S. 1). Der unmittelbare Adressat ist somit die Einzelschule mit ihren Lehrkräften als handelnde Akteure. Mit der Feststellung des Outputs über die Messung von Schülerleistungen bilden die Vergleichsarbeiten ein ex-

ternes Instrument, welches die interne individuelle *Schulevaluation* anregen und unterstützen soll (vgl. Kultusministerkonferenz, 2006). Mithilfe der Überprüfung, ob die Bildungsstandards in der eigenen Schule erreicht worden sind, können wiederum Rückschlüsse auf die Lernprozesse hinsichtlich ihrer Wirksamkeit gezogen werden, so dass an dieser Stelle eine *Qualitätssicherung* stattfindet (vgl. Heymann, 2005a, S. 24). Die Ergebnisse aus den Tests sollen nach Aussage der Landesinstitute somit zusätzliche, jedoch nicht die alleinigen Indikatoren bezüglich des Erfolgs von schulischen und unterrichtlichen Innovationen und Maßnahmen zur Verfügung stellen (vgl. Klieme, et al., 2007, S. 99; Peek, Pallack, Dobbelstein, Fleischer, & Leutner, 2006, S. 222). Wissenschaftlich betrachtet ist der Rückschluss eines Testergebnisses auf die Effektivität von Unterrichtsprozessen jedoch problematisch, da die Vergleichsarbeiten ohne Kenntnis der individuellen Unterrichtspraxis konzipiert werden. Eine Diskussion dieses Einwandes erfolgt im Abschnitt 4.4.1. Im Rahmen der Schulevaluation dienen die Vergleichsarbeiten jedoch nicht nur der Qualitätssicherung, sondern zugleich der *Weiterentwicklung schulischer Qualität*, indem aufbauend auf der Auswertung der Tests weitere schulische Konzeptionen initiiert werden können (vgl. Bonsen, Büchter, & Peek, 2006, S. 135). Die Vergleichsarbeiten verfolgen laut Aussage der EMSE somit einerseits summative Funktionen, die mit den Stichwörtern Messen, Prüfen und Sichern zusammengefasst werden können, und andererseits auch formative Aspekte, in Form einer Anregung von Entwicklung, Förderung und Unterstützung (vgl. Netzwerk Empiriegestützte Schulentwicklung, 2006, S. 1). Bei dieser Aussage wird erneut die Zwiespältigkeit der Tests hinsichtlich ihrer zu erfüllenden Funktionen deutlich.

Im Rahmen der summativen Testbestandteile liefern die Vergleichsarbeiten eine *Bestandsaufnahme* zu den vorhandenen Lernständen von Schulklassen und einzelnen Lerngruppen (vgl. Groß Ophoff, Koch, Hosenfeld, & Helmke, 2006, S. 21). Mittels der empirischen Validierung der Tests kann eine Verteilung der Ergebnisse auf Klassenebene – nicht auf Schülerebene – auf die einzelnen Kompetenzstufen vorgenommen werden (vgl. Höfer, Steffens, Diehl, Loleit, & Maier, 2009, S. 8). Hierfür ist es erforderlich, dass die Tests stark an den Bildungsstandards und den zugehörigen Kompetenzmodellen ausgerichtet sind. Zudem ist die Berücksichtigung psychometrischer Aspekte für die spätere kriteriale Interpretation des Ergebnisses von enormer Bedeutung (vgl. Klieme & Leutner, 2006, S. 877). Für die differenzierte Abbildung der Schülerleistung wird eine hinreichend große Anzahl von Aufgaben benötigt, welche die diversen Kompetenzniveaus präzise abbilden sowie tatsächliche Fähig- und Fertigkeiten anstelle von trägem Wissen überprüfen. Zudem kann mittels der Darstellung der Kompetenzverteilung einer Klasse der Lernstand in Hinblick auf die mittelfristig zu erreichenden Schulabschlüsse bzw. auf den gelingenden Übertritt von der Primar- zur Se-

kundarstufe abgeschätzt werden. Die Vergleichsarbeiten stellen hierbei eine ergänzende Leistungsbeurteilung für die Lehrkraft dar. Nach Ansicht von Orth (vgl. 2001, S. 217) verfolgen die Vergleichsarbeiten gleichzeitig das Ziel, aufgrund der empirischen Testgrundlage eine objektive Lernstandsermittlung anzubieten, welche die Schwächen alltäglicher Lehrerbeurteilungen korrigiert.

Des Weiteren ermöglichen die Vergleichsarbeiten eine *Verortung* der Schülerleistungen über das Aufzeigen vergleichender Perspektiven. Diese Verortung erfolgt einerseits kriterial über das Referenzsystem Bildungsstandards, da die Lernstände an den einheitlich definierten Anforderungen gemessen werden. Andererseits kann die Schulklasse aufgrund der Ermittlung von Vergleichswerten von Lerngruppen mit ähnlichen Ausgangsbedingungen, beispielsweise in Bezug auf die Zusammensetzung der Schülerschaft, sozial verortet werden (vgl. Köller, 2007, S. 101; Netzwerk Empiriegestützte Schulentwicklung, 2006, S. 1).

Durch diesen Vergleich der Testergebnisse mit Referenzklassen nehmen die Leistungstests die Funktion einer *Rechenschaftslegung* ein. Die Schulen und die einzelnen Lehrkräfte übernehmen Verantwortung für die Erfüllung des an sie gestellten Bildungsauftrages und geben somit implizit ein Zeugnis über ihre geleistete Arbeit ab (vgl. Tresch, 2007, S. 45). Diese Rechenschaftslegung kann gegenüber der Schulverwaltung, den Schülern, den Eltern und auch in reflexiver Weise gegenüber sich selbst erfolgen. Heymann (vgl. 2005b, S. 8) verbindet mit diesem Aspekt ebenso die Hoffnung, dass die Pflicht zur Rechenschaft mit einer erhöhten Motivation und einem größerem Engagement der Lehrpersonen einhergehe. Dies sei jedoch angezweifelt, bis entsprechende Ergebnisse aus wissenschaftlichen Untersuchungen zu diesem Aspekt vorliegen.

Als formativer Bestandteil der Vergleichsarbeiten wurde die Qualitätsentwicklung angesprochen. Diese soll insbesondere auf der operativen Ebene im Bereich der Unterrichtsentwicklung zum Tragen kommen (vgl. Tresch, 2007, S. 47). Die Ergebnisse fungieren in diesem Sinne als eine extern gelieferte Grundlage für nachfolgende pädagogische und fachdidaktische Handlungen, so dass weiterführende Lernprozesse angestoßen werden können. Verbunden ist damit eine *Innovationsfunktion* der Vergleichsarbeiten, da mithilfe kompetenzorientierter Aufgabenstellungen Impulse für einen anregenden und schülerorientierten Unterricht gesetzt werden (vgl. Heymann, 2005a, S. 25). Das Landesinstitut des Bundeslandes Hessen beabsichtigt mit der neuartigen Testkonzeption in diesem Zusammenhang eine allmähliche Veränderung der Aufgabenkultur (vgl. Hessisches Kultusministerium, 2009, S. 11). Das ehemals verantwortliche Institut für VERA in der vierten Klassenstufe geht diesbezüglich einen Schritt weiter, indem die Vergleichsarbeiten eine stärkere ergebnisorientierte Perspektive im Unterricht befördern sollen (vgl. Groß Ophoff,

Koch, Hosenfeld, & Helmke, 2006, S. 22). Dass mit diesen Aspekten zugleich Risiken impliziert sein können, wird in Abschnitt 4.4.2 thematisiert.

Ein weiterer Aspekt der Unterrichtsentwicklung ist die Funktion des *Diagnostizierens und Förderns*, welche den Bereich der Individualdiagnostik berührt. Mittels der Testergebnisse erhalten die Lehrpersonen zusätzliche Informationen über die Stärken und Schwächen ihrer Lerngruppe, so dass darauf aufbauend eine didaktische Schwerpunktsetzung im Unterricht erleichtert wird (vgl. Heymann, 2005b, S. 8). Auf diese Weise sollen Anforderungsprofile und spezifische passgenaue Pläne zur weiteren Förderung von den Lehrkräften aufgestellt werden (vgl. Höfer, Steffens, Diehl, Loleit, & Maier, 2009, S. 8). Dabei sind die Vergleichsarbeiten als ein Instrument zur Förderung anstatt zur Selektion zu verwenden.

Die Funktion der Förderung ist im bildungspolitischen und erziehungswissenschaftlichen Diskurs unumstritten; debattiert wird hingegen, ob die Vergleichsarbeiten für eine Individualdiagnostik auf der Ebene des einzelnen Schülers geeignet sind oder lediglich Aussagen auf Klassenbasis zulassen. Beispielsweise betonen die Landesinstitute von Hessen und Nordrhein-Westfalen sowie auch die EMSE, dass die Vergleichsarbeiten ein Planungsinstrument auf Klassenebene darstellen und sich daher nicht originär zur Individualdiagnostik eignen (vgl. ebd., S. 8; Netzwerk Empiriegestützte Schulentwicklung, 2006, S. 3; Peek, Pallack, Dobbelstein, Fleischer, & Leutner, 2006, S. 222). Demgegenüber lassen sich jedoch auch Äußerungen von Wissenschaftlern und wiederum der EMSE anführen, welche ausdrücklich eine individuelle Förderung einzelner Schüler in Verbindung mit den Tests sehen (vgl. Criblez, Oelkers, Reusser, Berner, Halbheer, & Huber, 2009, S. 46; Netzwerk Empiriegestützte Schulentwicklung, 2008, S. 3; Regenbrecht, 2005, S. 21). Festzuhalten bleibt daher, dass die Tests grundsätzlich in Hinblick auf eine Auswertung auf Klassenebene konzipiert werden. Für den Fall, dass die Ergebnisse zusätzlich individualdiagnostische Hinweise liefern, wäre dies ein positiver Nebeneffekt. Inwiefern dieser tatsächlich eintritt, ist mithilfe empirischer Untersuchungen zu überprüfen.

Da die Schulentwicklung aufgrund der Vergleichsarbeiten angetrieben werden soll, werden die Tests des Weiteren mit der Funktion einer *Weiterentwicklung in der Lehrerprofessionalität* verknüpft. Mit der Rückmeldung der Testergebnisse erhält die Lehrperson eine indirekte Rückmeldung zu ihrer eigenen pädagogischen Arbeit, so dass die *Selbstreflexion* hinsichtlich des unterrichtlichen Erfolgs angeregt wird (vgl. Regenbrecht, 2005, S. 21). Eine besondere Rolle wird hierbei der Verbesserung der *diagnostischen Kompetenz* zugeschrieben, da eigene Leistungseinschätzungen mit der objektiven externen Erhebung abgeglichen werden können (vgl. Criblez, Oelkers, Reusser, Berner, Halbheer, & Huber, 2009, S. 46). Diese Ergänzung des Lehrerurteils kann insbesondere bei der Grundschulempfehlung für den Über-

tritt in die Sekundarstufe herangezogen werden (vgl. Stähling, 2005, S. 213). Die Lehrkräfte werden mittels der Vergleichsarbeiten dahingehend sensibilisiert, dass sie „sowohl die relevanten Indikatoren der Leistungsfähigkeit erkennen als auch die notwendigen diagnostischen Methoden erwerben, um sich ein zutreffendes Urteil vom aktuellen Leistungsvermögen der Schüler bilden zu können“ (Levin, 2009, S. 19). Hinzu kommt eine verstärkte *Orientierung an den Bildungsstandards*, was im Wirkungskreislauf wiederum die Durchführung eines *kompetenzorientierten Unterrichts* befördern soll (vgl. Netzwerk Empiriegestützte Schulentwicklung, 2006, S. 1). Die Tests leisten in diesem Sinne den Beitrag, nicht bloßes Faktenwissen, sondern Kompetenzen in Form von Fähig- und Fertigkeiten zu komplexen Anforderungssituationen zu überprüfen (vgl. Ballasch, 2009, S. 302). Neben der Förderung dieser diagnostischen, didaktischen und methodischen Kompetenzen tritt der *kompetente Umgang mit den Informationsmedien* hinzu, da die Dateneingabe und Rückmeldung der Ergebnisse auf statistischer Basis über das Internet übertragen werden (vgl. Groß Ophoff, Koch, Hosenfeld, & Helmke, 2006, S. 21).

Im Rahmen der Schulentwicklung forcieren die Vergleichsarbeiten eine *intensivierte Fachgruppen- und Fachkonferenzarbeit*. Die Ergebnisse sollen in enger Kooperation der Kollegen untereinander analysiert und bewertet werden, so dass darauf aufbauend Handlungsschritte entwickelt werden können (vgl. Ballasch, 2009, S. 302). Dies befördert zugleich einen Austausch über grundlegende fachliche und pädagogische Fragen zu Unterricht und Schule (vgl. Orth, 2001, S. 206), wie beispielsweise über die Entwicklung eines schulinternen Lerncurriculums.

Nachdem die intendierten Funktionen von Vergleichsarbeiten für die Bereiche Schul- und Unterrichtsentwicklung beleuchtet worden sind, folgt nun die Darstellung des Zusammenhangs zwischen den Leistungstests und dem Bildungsmonitoring. Zum einen werden mithilfe der Vergleichsarbeiten zusätzliche Informationen über die Lernstände der Schüler eines Bundeslandes gewonnen. Diese Daten können als *Bestandsaufnahme* in ein Planungs- und Steuerungswissen auf Systemebene generiert werden. Die Vergleichsarbeiten können somit als ein weiteres Instrument der länderbezogenen Maßnahmen zur Qualitätssicherung und -entwicklung betrachtet werden. Das Konzept eines länderübergreifenden Austausches der Schülerergebnisse wird zunächst nicht verfolgt. Zusätzlich kann durch die Analyse der schulbezogenen Ergebnisse *Entwicklungs- und Förderbedarf in Hinblick auf schulische Unterstützungssysteme* festgestellt werden (vgl. Levin, 2009, S. 27).

Eine weitere Funktion kommt den Vergleichsarbeiten als unterstützendes Instrument bei der *Implementierung der Kerncurricula und Bildungsstandards* zu (vgl. Stähling, 2005, S. 213). Durch die kompetenzorientierte Gestaltung der Tests wird die Lehrerschaft unmittel-

bar mit der praktischen Anwendung der Standards konfrontiert. Über die indirekt „erzwungene“ Auseinandersetzung mit den neuen Steuerungsmechanismen soll die Implementierung beschleunigt und die Akzeptanz seitens der Schulakteure erhöht werden. Die enge Anlehnung der Vergleichsarbeiten an die Kompetenzmodelle soll den zusätzlichen Effekt bewirken, dass die zu bewältigenden Anforderungen, wie sie mit den Standards determiniert wurden, *transparenter* erscheinen (vgl. Orth, 2001, S. 206 f.). Im Zuge einer erhöhten Vergleichbarkeit werden die Verbindungen zwischen zu absolvierenden Lern- und Testgegenständen ersichtlich, so dass nach van Ackeren die Vergleichsarbeiten zugleich als „normorientierte Impulsgeber“ (2003, S. 43) zu verstehen sind.

Als eine letzte Funktion der Vergleichsarbeiten im Rahmen des Bildungsmonitorings ist anzuführen, dass die Tests als ein empirisches *Validierungsinstrument für die entwickelten Kompetenzmodelle* herangezogen werden können (vgl. Dobbelstein & Peek, 2004, S. 16).

Die Mehrdeutigkeit hinsichtlich der Funktion von Vergleichsarbeiten wirft gehäuft offene Fragen bezüglich ihrer Wirksamkeit auf (vgl. Altrichter, 2010, S. 228; Maier, 2010b, S. 48). Bereits Klieme, et al. (vgl. 2007, S. 87) rieten in ihrer Expertise von einer Überlagerung verschiedener Funktionen innerhalb eines Testmodells ab, sondern empfahlen eine deutliche Trennung. Die vorliegende Arbeit verfolgt daher das Ziel, die Konstruktion und Wirksamkeit der Vergleichsarbeiten in Bezug auf ihre intendierten Funktionen vertieft zu überprüfen.

4.3 Konstruktion der Vergleichsarbeiten

Die Vergleichsarbeiten durchlaufen von der Erstellung bis zu ihrer Auswertung einen komplexen Prozess, welcher in mehrere Phasen eingeteilt werden kann: In der ersten Phase werden die Testaufgaben von speziell geschulten Personengruppen entwickelt und wissenschaftlich geprüft. Um mangelhafte Aufgaben aussortieren zu können, werden die Aufgaben in Phase 2 anhand einer repräsentativen Stichprobe pilotiert. Nach Auswertung der Ergebnisse können die endgültigen Testaufgaben durch eine Zuordnung zu den Kompetenzbereichen und -stufen skaliert werden (Phase 3). Die fertigen Testhefte werden an die Schulen weitergegeben, so dass die Testdurchführung an einem bundesweit einheitlich festgelegten Tag vorgenommen werden kann (Phase 4). Die Lehrkräfte korrigieren die Schülerleistungen und geben die Ergebnisse in einem Online-Portal ein. In Phase 5 erfolgt die Auswertung der Messungen über eine externe Instanz. Phase 6 besteht darauf basierend aus der statistischen und grafischen Aufbereitung der Testleistungen, welche als Rückmeldung an die Schulen weitergeleitet wird. Die Lehrpersonen und die Schulverwaltung haben anschließend die Aufgabe, die Ergebnisse für die Entwicklung und Umsetzung von Maß-

nahmen zur Steigerung der Unterrichts- und Schulqualität zu nutzen (Phase 7). Zugleich wird eine prozessbegleitende Evaluation der Vergleichsarbeiten und ihrer Aufgaben durch das jeweils zuständige Landesinstitut und das IQB vorgenommen (Phase 8). Darauf aufbauend startet unmittelbar die Entwicklung des nächsten Testdurchlaufes. Auf diese Weise wird ein Arbeitskreislauf generiert. Die Abbildung 5 fasst die vorgestellten Stadien grafisch zusammen, so dass ein Gesamtüberblick gewonnen werden kann. Eine detaillierte Beschreibung und Analyse der einzelnen Phasen erfolgt in den folgenden Abschnitten.

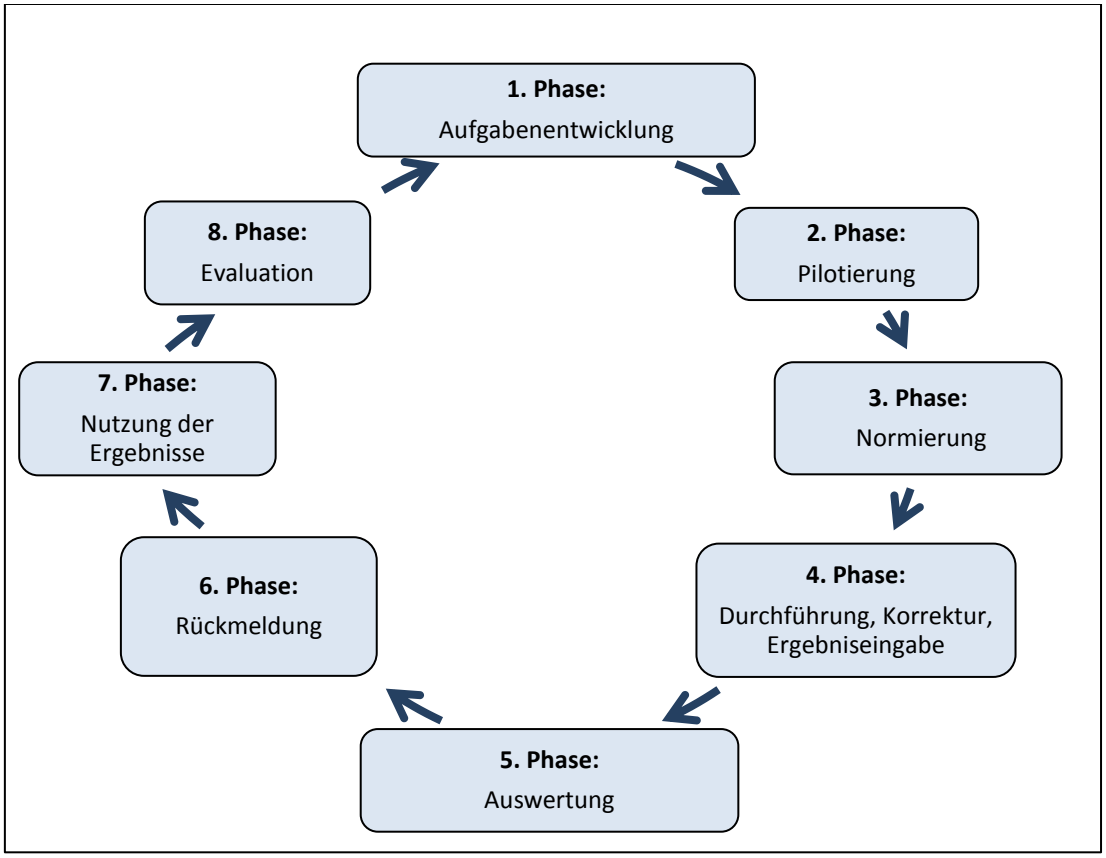

Abbildung 5: Phasenmodell eines Testdurchlaufs

4.3.1 Aufgabenentwicklung

4.3.1.1 Methodische Vorgehensweise und Zusammensetzung der Entwicklerteams

Die Vergleichsarbeiten werden mit dem Ziel eingesetzt, die erworbenen Fähig- und Fertigkeiten der Schüler zu erfassen. Folglich stellen die Tests bei ihrer Konzeption eine Operationalisierung der Kompetenzmodelle dar, da sich die Testinhalte sowohl den Kompetenzbereichen als auch den Anforderungsstufen zuordnen lassen müssen. Hierfür ist es erforder-

lich, eine große Zahl an Items zu entwickeln, so dass ein breites Spektrum der Schülerfähigkeiten abgebildet und erfasst werden kann. Ein Item wird nach Rost (vgl. 1996, S. 60) als die kleinste Beobachtungseinheit in einer Leistungsmessung verstanden. Beispielsweise wäre in einer Aufgabe mit a-, b-, c-Untergliederung der Aufgabenteil a) ein Item. Daher kann eine Aufgabe aus mehreren Items bestehen. Allgemein setzt sich ein Item aus den zwei Komponenten Itemstamm und Antwortformat zusammen. Der Itemstamm ist hierbei der Aufgabentext, in dem das situative Problem und die dazugehörige Aufgabenstellung vorgestellt werden. Bei diesem Aufgabenbestandteil ist eine transparente Anbindung an die Bildungsstandards herzustellen. Im Antwortformat wird demgegenüber die schriftlich fixierte Lösung des Getesteten aufgenommen (zu den verschiedenen Antwortformaten vgl. Abschnitt 4.3.1.3).

Um die Kompetenzmodelle vielfältig wiederspiegeln zu können, wird ein großer Itempool benötigt. Konkret bedeutet dies, dass nach Nennung des IQB bis zu 500 Items pro Fach bzw. sogar pro Kompetenzbereich konstruiert werden müssen (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2007, S. 20).

Für das Erreichen einer möglichst hohen Qualität der Vergleichsarbeiten werden die Aufgaben basierend auf den Grundzügen der klassischen und probabilistischen Testtheorie von speziell geschulten Personengruppen konzipiert und wissenschaftlich erprobt (vgl. Netzwerk Empiriegestützte Schulentwicklung, 2006, S. 2). Hierbei findet eine interdisziplinäre und länderübergreifende Kooperation zwischen erfahrenden Lehrkräften, Fachdidaktikern aus Hochschulen und Wissenschaftlern der empirischen Bildungsforschung aus dem IQB statt (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2007, S. 20). Die Federführung der Testentwicklung für VERA 3 und VERA 8 obliegt dem IQB, welches für jedes Fach eigene Arbeitsstrukturen konzipiert hat. Bei VERA 6 bestehen die Arbeitsgruppen zur Testerstellung ebenfalls aus Lehrpersonen, Fachdidaktikern und Koordinatoren aus den sechs teilnehmenden Bundesländern. Die wissenschaftliche Beratung und Unterstützung erfolgt in diesem Fall durch die Universität Jena (vgl. Ramm, 2006, S. 4).

Über die methodische Vorgehensweise innerhalb der Aufgabenentwicklungsgruppen gibt es wenig Kenntnis, da deren Arbeit in einem von außen unkontrolliertem Prozess verläuft (vgl. Meyerhöfer, 2005, S. 106). Jedes Mitglied der Erstellungsgruppe eines Fachs konzipiert entsprechend spezifischer Richtlinien eigene Items, die mit den zugehörigen Lösungen den Kollegen über ein internes Online-Portal zur Verfügung gestellt werden. Daran anknüpfend beginnt ein Diskussionsprozess, in welchem die Items nach diversen Gesichtspunkten, wie der Eindeutigkeit des Itemstamms oder der Lehrplanadäquatheit, untersucht und bewertet werden. Zusätzlich finden in regelmäßigen Abständen Zusammenkünfte der Entwickler-

gruppe statt, um die Prozesse zu koordinieren und generelle Absprachen zu treffen (vgl. Köller, 2010, S. 541 f.; Korngiebel, 2009, S. 46). Eine Methode zur Überprüfung der Items ist im Bereich der ersten Fremdsprache das Mappen. Hierbei wird der Aufgabentext in das Portal gestellt, verbunden mit der Aufforderung an die Kollegen, den Text hinsichtlich spezifischer Kriterien zu analysieren. Anschließend werden die einzelnen Ergebnisse untereinander verglichen. Wenn die gesammelten Informationen eine hohe Korrelation zueinander aufweisen, ist der Text für den Test brauchbar. Wenn dies jedoch nicht der Fall sein sollte, war entweder die Textauswahl ungeeignet oder der Itemstamm unpassend formuliert. Folglich setzt ein Interaktionsprozess ein, in dem das Item abgeändert oder der Text verworfen wird (vgl. Korngiebel, 2009, S. 46). Zusätzlich werden die Items durch Fachdidaktiker überprüft und abschließend von Wissenschaftlern bewertet. Somit ist jedes Item Objekt eines kollektiven Austausches zwischen Experten und kann mehrere Änderungen und Anpassungen erfahren. Dies ist mit einem hohen Arbeitsaufwand verbunden, da jeder Entwickler viele solcher Items begutachten muss (vgl. Meyerhöfer, 2005, S. 106).

Für eine passgenaue Anbindung an die Bildungsstandards werden bei der Erstellung der Vergleichsarbeiten sowohl fachdidaktische als auch psychometrische Aspekte berücksichtigt, was sich auch in der Interdisziplinarität der Arbeitsgruppen wiederspiegelt. Bei der Itemerstellung muss darauf geachtet werden, dass nicht nur didaktische Betrachtungen einfließen, sondern das Item auch hinsichtlich seiner Eignung als Messinstrument untersucht wird. „Das heißt, dass die didaktischen Gedanken in ein Fähigkeitskonstrukt umgesetzt werden müssen, welches dann operationalisiert werden muß, um das Vorhandensein der mit den didaktischen Gedanken verbundenen Fähigkeiten messen zu können. Ideen unterrichtlichen Handelns müssen also klar vom Problem der Fähigkeitenmessung getrennt werden“ (ebd., S. 106). Sicherlich muss eine didaktische Zielsetzung im Itemstamm erkennbar sein; letztlich ist ein Item jedoch nur dann brauchbar, wenn es später auch misst, was es messen soll. Daraus ergeben sich spezifische Anforderungen, welchen sowohl die einzelnen Items als auch der Test als Gesamtkonstrukt genügen müssen.

4.3.1.2 Aufgabenanforderungen

An einen „guten“ Test im Sinne eines qualitativ brauchbaren Messinstruments werden Anforderungen gestellt, die auf der wissenschaftlichen Testtheorie basieren. Differenziert wird zwischen den drei Gütekriterien *Validität*, *Reliabilität* und *Objektivität*.

Validität

Unter Validität versteht Rost (vgl. 1996, S. 31) den Gültigkeitsgrad des Tests bzw. die Aussagekraft des Testergebnisses bezüglich des ursprünglichen Messziels. Die Kernfrage bei der Untersuchung nach Validität lautet demnach: Misst der Test, was er messen soll? Beispielsweise wird von einem Item einer Vergleichsarbeit in Mathematik erwartet, dass es ausschließlich die mit der Aufgabe verbundene mathematische Kompetenz misst. Wenn allerdings der Itemstamm einen enorm hohen Textanteil aufweist, wird in dieser Aufgabe unweigerlich das Leseverständnis parallel geprüft. Dies würde in strenger Auslegung nicht der Anforderung von Validität genügen (vgl. Granzer, 2006, S. 19). Natürlich lassen sich Elemente der Lesekompetenz in keiner Textaufgabe verhindern; die zu messende Fähigkeit sollte aber klar im Zentrum des Items stehen.

Die Validität als Testgütekriterium kann nochmals differenziert werden (vgl. Bortz & Döring, 2002, S. 198 ff.; Rost, 1996, S. 32 f.):

- *Inhaltsvalidität*

  Bei der Inhaltsvalidität werden die Items darauf überprüft, ob sie die spezifische Kompetenz hinreichend erfassen. Vorgenommen wird dies durch die subjektive Einschätzung der Experten aus der Gruppe der Testersteller.

- *Kriteriumsvalidität*

  Für die Ermittlung der Kriteriumsvalidität werden Außenkriterien verwendet, deren Übereinstimmung mit dem Testergebnis untersucht wird. Als Außenkriterien können entweder ein zweiter vergleichbarer Test (externe Validität) oder eine vorhersagende Einschätzung zum Antwortverhalten (interne Validität) fungieren. Letzteres fußt auf präexperimentellen Hypothesen oder auf Theorien der Testersteller.

- *Konstruktvalidität*

  Die Konstruktvalidität stellt gewissermaßen eine Integration von Inhalts- und Kriteriumsvalidität bezüglich des Testgesamtkonstrukts dar, indem Annahmen getroffen werden, deren Zusammenhänge mittels Korrelationen überprüft werden.

Reliabilität

Das zweite Anforderungsmerkmal für Testaufgaben, die Reliabilität, lässt Aussagen über die Zuverlässigkeit und die Messgenauigkeit zu. Im Zentrum steht hierbei das Ausmaß, wie genau das Item etwas misst (vgl. Rost, 1996, S. 31). Betrachtet wird daher der Messfehlerbereich: Je kleiner dieser ist, desto höher ist die Reliabilität (vgl. Bortz & Döring, 2002, S.

195). Grundsätzlich existieren drei Vorgehensweisen zur Überprüfung der Reliabilität (vgl. ebd., S. 196 ff.; Schwippert, 2005a, S. 16):

- *Paralleltest-Reliabilität*

  Bei dieser Methode werden zwei inhaltlich sehr ähnliche Tests beim gleichen Probanden eingesetzt. Der Grad der Ergebnisübereinstimmung gibt Auskunft über die Zuverlässigkeit des Tests. Allerdings ist eine exakte Bestimmung der Reliabilität auf diese Weise nicht möglich, da es sich nur um ähnliche, aber nicht um identische Testformate handelt.

- *Testhalbierungsreliabilität*

  Im Rahmen der Testhalbierungsreliabilität werden die Messungen aus zwei Testhälften miteinander verrechnet, was jedoch nur bei äußerst umfangreichen Tests mit einer hohen Anzahl an Items praktizierbar ist. Die Zeitdauer zur Bearbeitung der Vergleichsarbeiten umfasst hingegen maximal achtzig Minuten, so dass jeweils nur eine begrenzte Itemmenge geprüft werden kann. Die Testhalbierungsreliabilität ist somit bei dem hier betrachteten Testkonzept eher als zweitrangig einzustufen.

- *Re-Test-Reliabilität*

  Bei dieser Verfahrensweise wird der Test unter identischen Bedingungen bei der gleichen Testperson zu einem zweiten Messzeitpunkt wiederholt. Sind die Ergebnisse zueinander hinreichend äquivalent, gelten die Tests als reliabel. Eine absolute Übereinstimmung wird jedoch auch hier nicht zu erreichen sein, da Erinnerungseffekte auftreten und sich Situationsveränderungen, verbunden mit Faktoren wie Konzentrationsverlust, nicht vermeiden lassen.

Objektivität

Das dritte Gütekriterium, die Objektivität, charakterisiert die Unabhängigkeit des Tests von verschiedenen Merkmalen. Rost (vgl. 1996, S. 37 f.) unterscheidet fünf Arten der Objektivität voneinander, die jedoch nicht alle primär der Phase der Aufgabenentwicklung zuzuordnen sind. Der Vollständigkeit halber werden sie aber an dieser Stelle gebündelt betrachtet:

- *Durchführungsobjektivität*

  Die Durchführungsobjektivität sagt aus, dass die Einweisung und Bearbeitung des Tests personenunabhängig sind. Bezogen auf die Vergleichsarbeiten sollte es zum Beispiel unerheblich sein, welche Lehrkraft die Aufsicht führt, da ein Item objektiv ist, wenn es ohne zusätzliche individuelle Hilfestellungen durch die Lehrpersonen von den Schülern bearbeitet werden kann.

- *Auswertungsobjektivität*

  Die Auswertungsobjektivität umfasst die Unabhängigkeit des Messergebnisses von der Auswertungsperson. Eine Zweitkorrektur sollte daher das gleiche Ergebnis ermitteln wie die Erstkorrektur.

- *Interpretationsobjektivität*

  Hiermit wird die Unabhängigkeit von der Person, welche die Messdaten anschließend interpretiert, ausgedrückt.

- *Situationsobjektivität*

  Im Sinne der Situationsobjektivität sollte das Testergebnis in keinem Zusammenhang zu der Situation während der Testdurchführung stehen. Extreme Bedingungen werden hierbei ausgeschlossen.

- *Spezifische Objektivität*

  Spezifische Objektivität bedeutet eine Unabhängigkeit von der Aufgabenauswahl. Die Grundlage hierfür bildet ein erstellter Pool an Items gleichen Typs und identischer Zielsetzung. Welche von diesen Items letztendlich im Test eingesetzt werden, steht ohne Beziehung zu dem Leistungsergebnis und dessen Interpretation. Wenn beispielsweise ein Schüler in der Vergleichsarbeit für die erste Fremdsprache eine Aufgabe lösen kann, die dem GER-Niveau A2 (vgl. Abschnitt 2.3.2.2) zugeordnet ist, sollte der Proband im Sinne der spezifischen Objektivität bei einem weiteren A2-Item die gleiche Fähigkeit darlegen können. Aus diesem Grund ist es unverzichtbar, dass die Testersteller ein breites Aufgabenspektrum mit mehreren Items für jeden Kompetenzbereich und die zugehörigen Kompetenzstufen entwickeln.

Bei Betrachtung dieser drei Gütekriterien können Wirkungszusammenhänge festgestellt werden: Die Objektivität bildet die notwendige Voraussetzung für die Reliabilität, da ein nicht-objektiver Test unweigerlich keine hohe Messgenauigkeit aufweisen kann (vgl. ebd., S. 39). Die Reliabilität bildet wiederum die Grundlage für die Validität. Arnold führt als weitere Gütekriterien die Vergleichbarkeit mit anderen Testverfahren, Ökonomie in Hinblick auf die aufzuwendenden Ressourcen, Nützlichkeit bezüglich des Testzwecks sowie Fairness an (vgl. Arnold, 2001). Daneben müssen die Aufgaben den aktuellen Forschungsergebnissen der Fachwissenschaft, Entwicklungspsychologie und Pädagogik entsprechen.

An die Items der Vergleichsarbeiten werden insbesondere inhaltliche Anforderungen gestellt. Die Messungen verfolgen die Zielsetzung, sowohl die Kompetenzstruktur- als auch die Kompetenzstufenmodelle über die Aufgaben zu operationalisieren (vgl. Klieme & Leutner,

2006, S. 877). In der Praxis bedeutet dies, dass sich jedes Item genau zu einem Kompetenzbereich und einer Anforderungsstufe zuordnen lassen muss. Hier greift erneut das Merkmal der Validität, indem mit einem Item nicht mehrere Fähig- und Fertigkeiten gleichzeitig gemessen werden sollten. Dass dies in der praktischen Anwendung nicht immer möglich ist, zeigt der folgende Ausschnitt eines Items aus VERA 8 Mathematik, 2009 (vgl. Abbildung 6).

Aufgabe 6: Quiz

Aufgabe 6.1: Quiz
Birgit nahm an einem Quiz teil, bei dem sie insgesamt 18 Fragen zu beantworten hatte. Für jede richtige Antwort erhielt sie einen Pluspunkt, für jede falsche Antwort einen Minuspunkt. Am Ende des Quiz hatte sie acht Pluspunkte.
Wie viele Fragen hatte Birgit insgesamt richtig beantwortet?

Abbildung 6: Beispiel für eine Zuordnung eines Items zu den Kompetenzbereichen (vgl. Institut für Qualitätsentwicklung, 2009, S. 6)

Generell erfordert diese Aufgabe kombinatorische Überlegungen, so dass sie klar der Leitidee *Zahl* (L1) zugeordnet werden kann. Für die Lösung des Items 6.1 muss der Schüler sowohl eine Strategie entwickeln als auch einfache Rechenoperationen ausführen. Daher bildet dieses Item die Kompetenz *Probleme mathematisch lösen* (K2) sowie die Kompetenz *mit Mathematik symbolisch/ formal/ technisch umgehen* (K5) ab (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2009a, S. 26 f.). Die Lösung dieses Aufgabenbestandteils fordert somit mehrere Fähigkeiten gleichzeitig, was dem Anspruch nach Eindeutigkeit entgegensteht. Folglich müsste für die Auswertung der einzelnen Kompetenzbereiche das Item doppelt herangezogen werden. Für eine eindeutige Zuordnung ist es bedeutsam, dass die Items genau das erfassen, was für jenes angesprochene Niveau charakteristisch ist. Es ergibt sich die Notwendigkeit, bereits während der Konstruktion eines Items festzuhalten, welche Wissensbestandteile und Prozesse zur Bewältigung erforderlich sind (vgl. Klieme, et al., 2007, S. 86).

Des Weiteren bedarf es einer inhaltlichen Sinnhaftigkeit der Aufgaben (vgl. van den Heuvel-Panhuizen, 2006, S. 16). Indem sie reale Lebenssituationen widerspiegeln, wird die Motivation der Schüler zur Lösung der Aufgabe erhöht und die Lernenden erkennen einen Nutzen des Items. Die Testaufgaben sollten in ihrer Formulierung innovative Impulse für die weitere Unterrichtsgestaltung aussenden. Aus diesem Grund wird in Form einer Handreichung die Zuordnung der Items zu den Kompetenzmodellen sichtbar gemacht und die didaktisch-methodische Begründung der Aufgaben erläutert. Dies bildet die Voraussetzung für die

spätere Interpretation in Hinblick auf die zukünftige Unterrichtsgestaltung und auf Förderhinweise (vgl. Schirp, 2006b, S. 432).

Zudem sei an dieser Stelle nochmals die Funktion der Vergleichsarbeiten erwähnt, eine neue innovative Aufgabenkultur im Rahmen eines kompetenzorientierten Unterrichts zu befördern. Dabei müssen Testaufgaben von Lernaufgaben klar voneinander abgegrenzt werden. Mittels Testaufgaben kann der aktuelle Leistungsstand in Form einer Kompetenzerhebung möglichst umfassend ermittelt werden. Hierzu haben sie den psychometrischen Anforderungen zu genügen und müssen in kurzer Zeit bearbeitbar sowie objektiv auswertbar sein (vgl. Speck-Hamdan, 2007, S. 5; Steinweg, 2007, S. 5). Die Testaufgaben sind somit lediglich auf messbare Merkmale ausgelegt, aber nicht für den Erwerb neuer Kompetenzen geeignet. Die im Unterricht eingesetzten Aufgabenstellungen sollten sich in ihrer Konzeption klar von den Testaufgaben unterscheiden, indem die Lernaufgaben relativ offen und auf komplexe Situationen bezogen konstruiert werden (vgl. Drieschner, 2009, S. 93). Mittels gehaltvoller Arbeitsaufträge sollten im Unterricht vorhandene Kompetenzen miteinander in neuartigen Situationen verknüpft werden bzw. neue Fähig- und Fertigkeiten ausgebildet werden. Bei nochmaliger Betrachtung des Modelles der Input- und Outputsteuerung können die Testaufgaben somit beim Output verortet werden, da sie die Ausprägungen der Kompetenzen zu einem bestimmten Zeitpunkt des Lernprozesses offenbaren. Lernaufgaben ermöglichen jedoch erst diesen Lernprozess, weil sie für den Erwerb von Fähig- und Fertigkeiten eingesetzt werden (vgl. Abbildung 7).

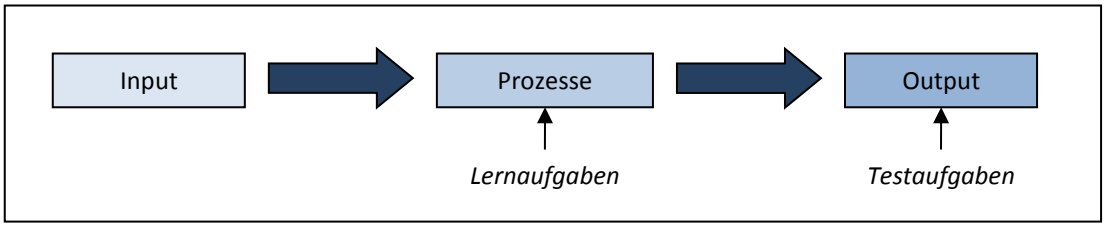

Abbildung 7:Verortung von Lern- und Testaufgaben in das Modell der Input- und Outputsteuerung

Testaufgaben können allerdings in abgewandelter Form als Lernaufgaben fungieren, so dass die Vergleichsarbeiten tatsächlich einen Impuls für die Aufgabenkultur in der Unterrichtspraxis zu leisten imstande sind. Dies erfordert jedoch ein hohes Maß an methodischer und didaktischer Professionalität sowie an Kreativität auf Seiten der Lehrpersonen.

4.3.1.3 Aufgabenformate

Um differenzierte Informationen zu gewinnen und die Vergleichsarbeiten einerseits ökonomisch und andererseits für Schüler anregend zu gestalten, entwickeln die Testersteller Items, die sich in drei verschiedene Formate untergliedern lassen: o*ffene*, *halboffene* und *geschlossene* Aufgaben. Nachfolgend wird jeder Typus hinsichtlich seines Nutzens für die Vergleichsarbeiten untersucht.

Offene Aufgaben

Offen gestellte Items verlangen freie, ausführliche Antworten oder Begründungen für eine Problemsituation. Im Itemstamm werden daher keine weiteren Vorgaben eingebaut (vgl. Bühner, 2006, S. 64). In den Vergleichsarbeiten sind trotz dieses Charakteristikums oftmals Hilfestellungen bezüglich der Vorgehensweise oder den erwarteten inhaltlichen Eckpunkten der Antwort vorgegeben, so dass die Lösung bereits im Vorfeld in die gewünschte Richtung gesteuert werden kann (vgl. Korngiebel, 2009, S. 53). Ob solche Items generell noch als offen einzustufen sind, ist diskutabel. In den Ausführungen dieser Arbeit werden sie zum offenen Format hinzugezählt, da sie dennoch ausführliche Lösungen verlangen.

Über offene Aufgaben kann beispielsweise ein vollständiger Rechenweg, eine Zeichnung, ein Kurzaufsatz oder die Äußerung der persönlichen Meinung zur Thematik eingefordert werden. Der Vorteil dieses Itemformates besteht darin, dass es sowohl eine freie Reproduktion von Wissen als auch einen kreativen Umgang mit der erworbenen Kompetenz erlaubt. Des Weiteren sind Zufallslösungen durch Raten ausgeschlossen (vgl. Bühner, 2006, S. 65). Offene Testitems sind in ihrer Konstruktion den Lernaufgaben ähnlich, die im regulären Unterricht Einsatz finden. Infolgedessen wird ein kompetenzorientierter Unterricht eine Konzentration auf die Verwendung von offenen Aufgaben verlangen (vgl. Lankes, 2006, S. 23).

Der Nachteil von offenen Itemformaten liegt im hohen Arbeitsaufwand. Baumert, et al. (vgl. 2000, S. 102) argumentieren, dass die Aufnahme offener Items in Tests wohl überlegt sein sollte, da die zusätzlichen Auswertungs- und Kodierungskosten oftmals in keinem Verhältnis zum relativ geringen Informationsgewinn ständen. Im Gegensatz zu geschlossenen Items bleibt dem Auswerter die Entscheidung überlassen, wie viele Punkte der Schülerantwort angemessen sind (vgl. Rost, 1996, S. 61). Um die Korrektur zu erleichtern, wird bereits bei der Aufgabenentwicklung eine Signierung vorgenommen, in der mögliche Schülerantworten prognostiziert werden.

Halboffene Items

Bei halboffenen Items wird im Vergleich zu offen gestellten Testfragen nur eine kurze Lösung in Form eines Stichwortes, Rechenergebnisses, maximal eines Satzes gefordert. Dies ist mit einer größeren Sicherheit für den Schüler verbunden, da der Erwartungshorizont transparenter gestaltet ist. Die Ratewahrscheinlichkeit ist bei diesem Format ebenfalls relativ gering. Zugleich können komplexe Anforderungssituationen hergestellt werden. Trotzdem ist die Kreativität bei halboffenen Items bereits eingeschränkt, weil eine klare, eindeutige Lösung erwartet wird. Dies impliziert einen geringeren Aufwand bei der Signierung als bei offenen Testfragen. Des Weiteren kann der Domino-Effekt eintreten und eine Kettenreaktion hervorrufen: Wenn der Schüler beispielsweise ein Wort fehlerhaft ergänzt, treten oftmals Folgefehler auf (vgl. Bühner, 2006, S. 65).

Geschlossene Aufgaben

Bei geschlossen formulierten Items wird dem Schüler bereits eine Auswahl an möglichen Lösungen offeriert, unter denen er sich für die richtige Aussage entscheiden muss. Der Schüler ist aus diesem Grund an die vorgefertigten Antworten gebunden. Dieses Aufgabenformat kann nochmals dahingehend ausdifferenziert werden, ob die Antwortmöglichkeiten disjunkt zueinander sind, also sich gegenseitig ausschließen, oder ob mit ihnen alle Lösungsmöglichkeiten ausgeschöpft sind, sie folglich exhaustiv sind.

Bei den geschlossenen Formaten gibt es mehrere Varianten für die Gestaltung der Items. Zum einen können *Multiple-Choice-Aufgaben* als die klassischen Vertreter des geschlossenen Aufgabentyps konzipiert werden, bei denen aus mehr als zwei Antwortmöglichkeiten die richtige angekreuzt werden muss. Um die Lösungswahrscheinlichkeit durch Raten so gering wie möglich zu halten, werden die falschen Antwortkategorien, die sogenannten Distraktoren, möglichst intelligent und plausibel formuliert. Auf diese Weise ist die Lösung nicht offensichtlich und die Lehrperson kann bei der Korrektur eventuelle Fehlerquellen und halbrichtige Lösungen analysieren (vgl. Rost, 1996, S. 64 ff.). Mittels einer Itemkonstruktion, bei der mehrere Antworten zu der richtigen Lösung kombiniert werden müssen, reduziert sich die Wahrscheinlichkeit für Zufallstreffer nochmals (vgl. Bühner, 2006, S. 57 ff.). Als zweite Form geschlossener Items sind *Richtig-/ Falsch-Aufgaben* zu nennen. Hier stehen lediglich zwei Alternativen zur Verfügung, so dass die Ratewahrscheinlichkeit 50 Prozent beträgt. Obwohl mit der Messung dieser Itemart nur wenige Informationen ermittelt werden können, bieten sie den Vorteil, dass sie von den Schülern relativ schnell verstanden werden (vgl. ebd., S. 56 f.). Beim dritten Typ erfolgt die Lösung über *Zuordnungen* von Aussagen oder Symbolen. Für den Schüler kann hierbei allerdings das Problem einer

Kettenreaktion von falschen Lösungen auftreten. Viertens können die Items als *Umordnungsaufgaben* gestaltet werden, bei denen alle vorgegebenen Antworten bereits richtig sind, aber entsprechend der Fragestellung in eine andere Reihenfolge zu sortieren sind.
Mit geschlossenen Formaten kann keineswegs ausschließlich Faktenwissen überprüft werden. Allerdings ist es bei einigen Kompetenzbereichen, wie dem *Schreiben* in Deutsch, problematisch, die Fähig- und Fertigkeiten auf diese Weise zu messen. Daher muss je nach dem zu prüfenden Kompetenzbereich und der damit verbundenen Zielsetzung entschieden werden, aus welchem Itemformat der größtmögliche Nutzen zur Informationsgewinnung gezogen werden kann. Der ökonomische Vorteil geschlossener Aufgaben liegt darin, dass in kurzer Zeit relativ viele Items bearbeitet und mit wenig Aufwand ausgewertet werden können. Somit kann eine größere Zahl an Items in den Test eingebaut werden, was wiederum die Reliabilität erhöht (vgl. Tresch, 2007, S. 96). Weil der Vorgang der Signierung entfällt, könnte vermutet werden, dass mit geschlossenen Items im Vergleich zu offenen Formaten eine höhere Objektivität verbunden ist. Bei einer sorgfältigen und für alle Beteiligten transparent vorgenommenen Signierung trifft dies jedoch nicht zu. Mit einer steigenden Anzahl an geschlossenen Items in einem Test sinkt jedoch die Validität, da mittels der vorgegebenen Antwortkategorien nicht alle Reaktionsmöglichkeiten der Schüler ausgeschöpft werden (vgl. Rost, 1996, S. 63). Während in den Lösungen offener Items die Gedankengänge der Schüler sichtbar werden, vermögen geschlossene Formate dies nicht zu leisten.

4.3.2 Pilotierung

Nachdem die Aufgabenentwickler die Items erstellt haben und der interne Diskussionsprozess abgeschlossen ist, findet die Präpilotierung statt. Dies geschieht oftmals durch die Entwickler selbst, indem sie die Aufgaben in ihrem eigenen Unterricht testen. Über diese ökonomische Variante fallen problematische Items heraus, die anschließend in einem gemeinsamen Arbeits- und Kommunikationsprozess überarbeitet werden (vgl. Hessisches Kultusministerium, 2009, S. 13).
Darauffolgend werden die Aufgaben des erstellten Item-Pools pilotiert. Dieser Prozess ist mit drei Zielsetzungen verbunden: Erstens wird die Tragfähigkeit der Aufgabenformulierung überprüft, das heißt eine Einschätzung über die Eignung eines Items als Testaufgabe abgegeben. Des Weiteren kann mit Hilfe der Pilotierung ein relativ großer Ausschnitt der potentiellen Schülerlösungen dokumentiert werden, so dass es nach Aussage von Ehmke, et al. (vgl. 2006, S. 226 f.) möglich ist, Fehlerhäufigkeiten zu sammeln und anschließend zu analysieren. Drittens können Anhaltspunkte für die inhaltliche Gestaltung der Korrekturanwei-

sung und Handreichung gewonnen werden. Eine Pilotierung dient somit der Ermittlung von Güte und Eignung der Items (vgl. Hessisches Kultusministerium, 2009, S. 13).

Die Pilotierung der Aufgaben wird in einer repräsentativen Stichprobe an Schulen in jedem beteiligten Bundesland von externen Testleitern durchgeführt. Die Größe der Stichprobe bei den Vergleichsarbeiten beträgt zwischen 2.000 und 3.000 Schülern. Auf diese Weise wird sichergestellt, dass die Durchführungsbedingungen in den Schulen möglichst äquivalent zueinander sind und die Geheimhaltung der Aufgaben gewahrt bleibt. Der entwickelte Item-Pool besteht aus einer großen Anzahl an Items, so dass es nicht realisierbar wäre, wenn jeder Schüler alle Aufgaben bearbeiten müsste. Der Pilotierung liegt aus diesem Grund ein sogenanntes Multiple-Matrix-Design zugrunde, welches auch in Large-Scale-Assessments Anwendung findet. Hierbei werden Itembündel zu verschiedenen Testheften zusammengefasst, die an die Schüler zufällig oder nach einem festgelegten Rotationsprinzip verteilt werden. Jeder Proband bearbeitet somit einen Teilausschnitt. Um eine möglichst genaue Messanalyse anfertigen zu können, wird jedes Item in mindestens zwei Testhefte aufgenommen. Die Testhefte sind auf diese Weise durch einige identische Items miteinander verknüpft (vgl. Baumert, Bos, & Lehmann, 2000, S. 56 ff.). Jede Aufgabe wird mit dieser Vorgehensweise von etwa 400 Schülern bearbeitet.

Die Messergebnisse der Pilotierung werden ebenso wie die der späteren Vergleichsarbeit mithilfe der probabilistischen Testtheorie ausgewertet (vgl. Abschnitt 4.3.3). Anschließend folgt die Interpretation der Daten und deren Beurteilung bezüglich der Eignung als Testaufgabe. Die Steuergruppen des IQB zogen beispielsweise für mathematische Items, die 2005 durch das Deutsche PISA-Konsortium pilotiert wurden, das folgende vierstufige Bewertungsmodell heran (vgl. Ehmke, Leiß, Blum, & Prenzel, 2006, S. 226 ff.):

Das Item ist als Testaufgabe

1. voll geeignet,
2. mit leichten Modifikationen geeignet,
3. nur nach umfangreicher Überarbeitung geeignet,
4. nicht geeignet.

Jene Items, die der Kategorie *voll geeignet* entsprechen, können ohne weitere Analyseverfahren als Testaufgaben genutzt werden. Bei den Items der Kategorie 2 werden konkrete Verbesserungsvorschläge eingebaut, so dass sie nach nochmaliger Überprüfung ebenfalls weiterverwendet werden können. Über Items der dritten Kategorie wird individuell entschieden, ob sie nochmals überarbeitet oder aussortiert werden. Die den Kategorien 2 und 3 zugeordneten Items durchlaufen somit den Aufgabenentwicklungsprozess zum Teil und

die Pilotierungsphase vollständig ein zweites Mal. Die *nicht geeigneten* Aufgaben der Kategorie 4 werden sofort aussortiert und nicht weiter bearbeitet. Gründe für die Zuordnung eines Items zur vierten Stufe können zum Beispiel darin gesehen werden, dass der Schwierigkeitsgrad der Aufgabe zu hoch ist und das Item in dessen Konsequenz nur von sehr wenigen Schülern gelöst werden kann. Außerdem kann der Signier- und Kodierungsprozess des Items aus ökonomischer Sicht nicht mehr vertretbar sein. Da einige Aufgaben über die Pilotierung ausgemustert werden, sind die Testentwickler in der ersten Phase der Aufgabenentwicklung dazu verpflichtet, ein Überangebot an Items zu konstruieren.

Um zu ermitteln, in welcher Kategorie ein Item zu verorten ist, wird zum einen seine Lösungshäufigkeit betrachtet, welche sich zwischen 5 Prozent und 95 Prozent bewegen sollte. Extrem leichte oder schwere Items werden somit ausgeschlossen, da keine aussagekräftigen Informationen gewonnen werden können. Zum anderen dient die zu ermittelnde Trennschärfe einer Aufgabe als ein Kriterium, indem das Item zwischen stark und schwach ausgeprägten Kompetenzen unterscheiden können sollte. Dabei wird die Beantwortung eines einzelnen Items mit der Gesamtleistung im Test basierend auf der Annahme verglichen, dass ein Schüler, der beispielsweise ein schweres Item erfolgreich lösen kann, eine hohe Gesamtpunktzahl im Sinne einer großen Trennschärfe aufweisen wird (vgl. Neumann, Karius, Robitzsch, Behrens, Krelle, & Böhme, S. 57). Des Weiteren erfolgt eine Überprüfung der Auswertungsobjektivität. Hierzu werden zum Teil Doppelkodierungen der dargelegten Schülerlösungen vorgenommen (vgl. Fleischer, Spoden, Wirth, & Leutner, 2008, S. 199).

4.3.3 Skalierung

Jene Items, die den Prozess der Pilotierung positiv durchlaufen haben, werden anschließend von den Testerstellern mithilfe statistischer Verfahren skaliert. Dies bedeutet, dass jedes Item einem Kompetenzbereich sowie einer -stufe zugeordnet wird. Die Verknüpfung einer Aufgabe mit einem spezifischen Kompetenzbereich sollte möglichst bereits in der Phase der Aufgabenentwicklung hergestellt worden sein. Mittels der gewonnenen Daten aus der Pilotierung können die getroffenen Annahmen wissenschaftlich überprüft und untermauert werden. Die Skalierung kann des Weiteren Erkenntnisse bezüglich der Verteilung der Testergebnisse generieren (vgl. Rost, 1996, S. 40). Die Zuweisung zu Niveaustufen gestaltet sich demgegenüber komplizierter, da zunächst die Schwierigkeit eines jeden Items zu bestimmen ist.

Die Basis für die Skalierung der Vergleichsarbeiten bilden die in PISA verwendeten Skalen, denen wiederum die probabilistische Testtheorie mit dem zugehörigen Raschmodell zu-

grunde liegt (vgl. Schweitzer, 2007, S. 61). Die Grundzüge der testtheoretischen Modellannahme werden an dieser Stelle vorgestellt. Für tiefergreifende Ausführungen sei auf die Publikationen von Baumert, et al. (1999; 2000), Bühner (2006), Neubrand (2004), Rost (1996) verwiesen.

Während bei der klassischen Testtheorie über die von der Testperson erreichte Gesamtpunktzahl unmittelbar auf die Personenfähigkeit unter Berücksichtigung von Messfehlern geschlossen wird, ist es bei der probabilistischen Testtheorie, auch Item-Response-Theorie genannt, möglich, die Personenfähigkeit und die Aufgabenschwierigkeit auf einer gemeinsamen Skala differenziert abzubilden (vgl. Nachtigall & Kröhne, 2006, S. 63). Dies hat den Vorteil, dass die Fähigkeiten der Probanden auch bei der Bearbeitung unterschiedlicher Aufgaben, wie es bei der Pilotierung im Rahmen des Multiple-Matrix-Designs abläuft, gemeinsam dargestellt werden können (vgl. Baumert, Bos, & Lehmann, 2000, S. 60 f.). Das Ziel des Modells besteht darin, die unterschiedlichen Fähigkeitsausprägungen der Probanden inhaltlich zu definieren und später einer Position auf der Testskala zuzuordnen (vgl. Baumert, Bos, & Watermann, 1999, S. 41). Hierfür muss zunächst eine Unterscheidung zwischen manifesten und latenten Variablen erfolgen. Manifeste Variablen sind beobachtbar, wie die Reaktion eines Schülers auf ein Item oder die anschließend gemessene Gesamtpunktzahl des Schülers. Demgegenüber ist die Fähigkeit einer Person als eine latente Variable nicht beobachtbar. Die Kompetenzausprägung eines Schülers gemäß seinem Testergebnis wird folglich nur durch eine normorientierte Interpretation der Testleistung offenbar. Diese latente Variable kann nun über die probabilistische Testtheorie quantitativ skaliert werden (vgl. Baumert, Bos, & Lehmann, 2000, S. 61).

Den einfachsten Fall stellt hierbei das dichotome Rasch-Modell dar. Zur Bestimmung der Aufgabenschwierigkeit wird angenommen, dass sich die manifeste Reaktion auf ein Item in zwei Kategorien unterteilen lässt: In erfolgreich gelöst oder nicht gelöst. Folglich können nur die Werte 1 (gelöst) oder 0 (nicht gelöst) angenommen werden. Jede manifeste Reaktion resultiert dabei aus der individuellen Personenfähigkeit und der Itemschwierigkeit. In Bezug auf die Vergleichsarbeiten bedeutet dies, dass die Schülerantwort sowohl von seiner Kompetenzausprägung, als auch von der Schwierigkeit der Aufgabe abhängt. Es lässt sich daher folgende logarithmische Gleichung aufstellen (vgl. ebd., S. 61 f.):

$$\ln\left(\frac{p\,(x_{vi})}{1-p\,(x_{vi})}\right) = \theta_v + \vartheta_i$$

Die statistische Wahrscheinlichkeit $p\,(x_{vi})$, dass der Schüler v das Item i löst, setzt sich additiv zusammen aus seiner Fähigkeit θ_v und der Schwierigkeit des Items ϑ_i. Durch Umstellungen ergibt sich schließlich die folgende Gleichung für die Lösungswahrscheinlichkeit:

$$p\,(x_{vi} = 1) = \frac{e^{(\theta_v + \vartheta_i)}}{1 + e^{(\theta_v + \vartheta_i)}}$$

Über diese Funktion kann für jedes Item die Lösungswahrscheinlichkeit im Zusammenhang mit der dazu erforderlichen Fähigkeit in der sogenannten Item Characteristic Curve (ICC) grafisch dargestellt werden. Bei Betrachtung der ICCs in Abbildung 8 wird ersichtlich, dass die Grundannahme darauf beruht, dass die Wahrscheinlichkeit, ein Item erfolgreich zu bewältigen, bei steigender Personenfähigkeit zunimmt.

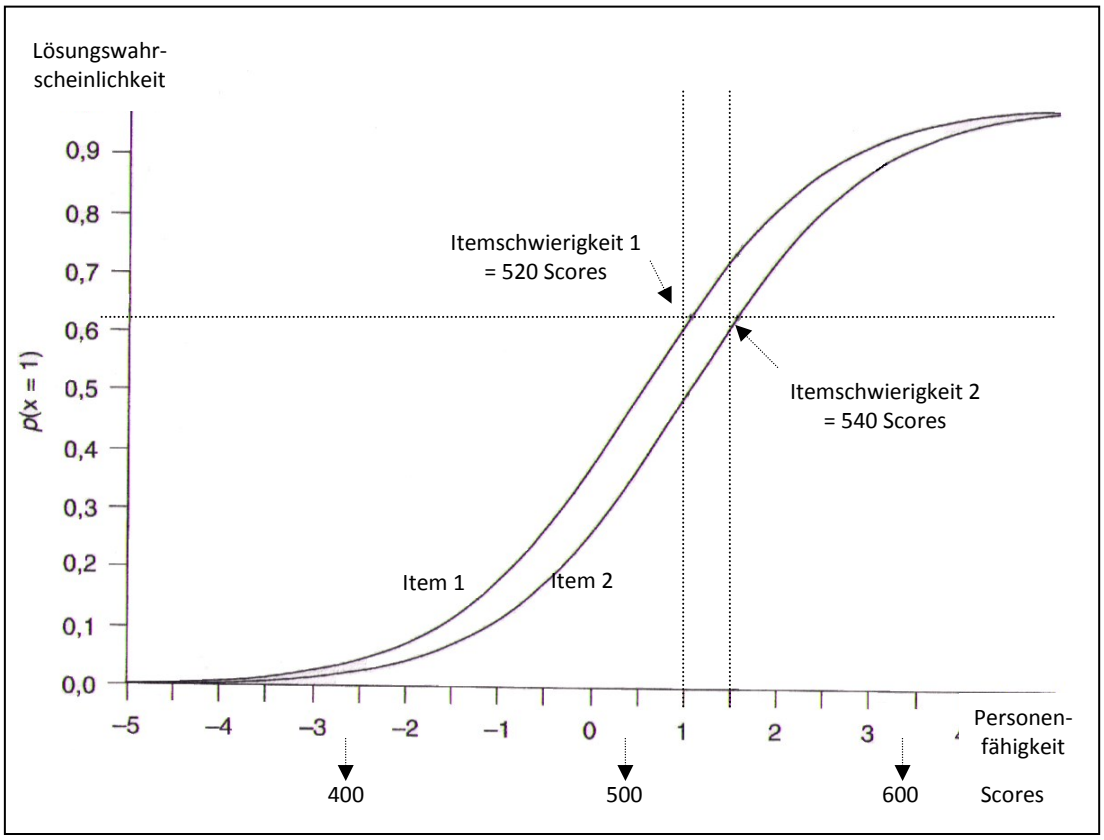

Abbildung 8:Item Characteristic Curve und die Bestimmung der Itemschwierigkeit (vgl. Baumert, Bos, & Lehmann, 2000, S. 63)

Mit Hilfe der Funktionsgleichung und der zugehörigen ICC kann daraufhin die jeweilige Aufgabenschwierigkeit berechnet werden, indem zunächst die über die Pilotierung ermittelte durchschnittliche Lösungshäufigkeit eines Items prozentual ausgedrückt wird. Zusätzlich werden die Werte auf der Fähigkeitsskala zu einer Score-Skala mit dem Mittelwert 500 mit einer Standardabweichung von 100 transformiert, so dass etwa zwei Drittel aller Schüler Werte zwischen 400 und 600 Scores erreichen. Die Schwierigkeit eines Items ist ursprünglich determiniert durch den Wendepunkt der ICC, bei welchem sich die Lösungswahrscheinlichkeit stets bei $p(x_{vi}) = 0{,}5$ befindet. Bei Large-Scale-Assessments wie PISA und DESI sind

jedoch besonders die Aufgaben von Interesse, die mit einer hinreichenden Sicherheit, das heißt mit einer Wahrscheinlich von 65 Prozent, gelöst werden (vgl. Baumert, Bos, & Watermann, 1999, S. 41). Köller (vgl. 2008b, S. 16) gibt hingegen an, dass den Items des IQB die Maßgabe von 62,5 Prozent zugrunde liegt, wie es auch in der Abbildung 8 übernommen wurde. Im obigen Beispiel beträgt die Schwierigkeit des Items 1 dementsprechend etwa 520 Scores, die des zweiten Items circa 540 Scores.

Die Skalierung von Items wird insgesamt mehrmals mit verschiedenen Schülergruppen, zum Beispiel differenziert nach Geschlecht oder Schulart, durchgeführt. Wenn die Ergebnisse hierbei große Unterschiede zueinander aufweisen, ist dies ein Indiz dafür, dass das Item nicht ausschließlich die beabsichtigte Kompetenz misst. Folglich müsste das Item von der weiteren Testung ausgeschlossen werden.

Das bereits in Abschnitt 2.3.2.2 vorgestellte Kompetenzstufenmodell kann mit den Score-Werten erweitert werden, wie in Abbildung 9 deutlich wird.

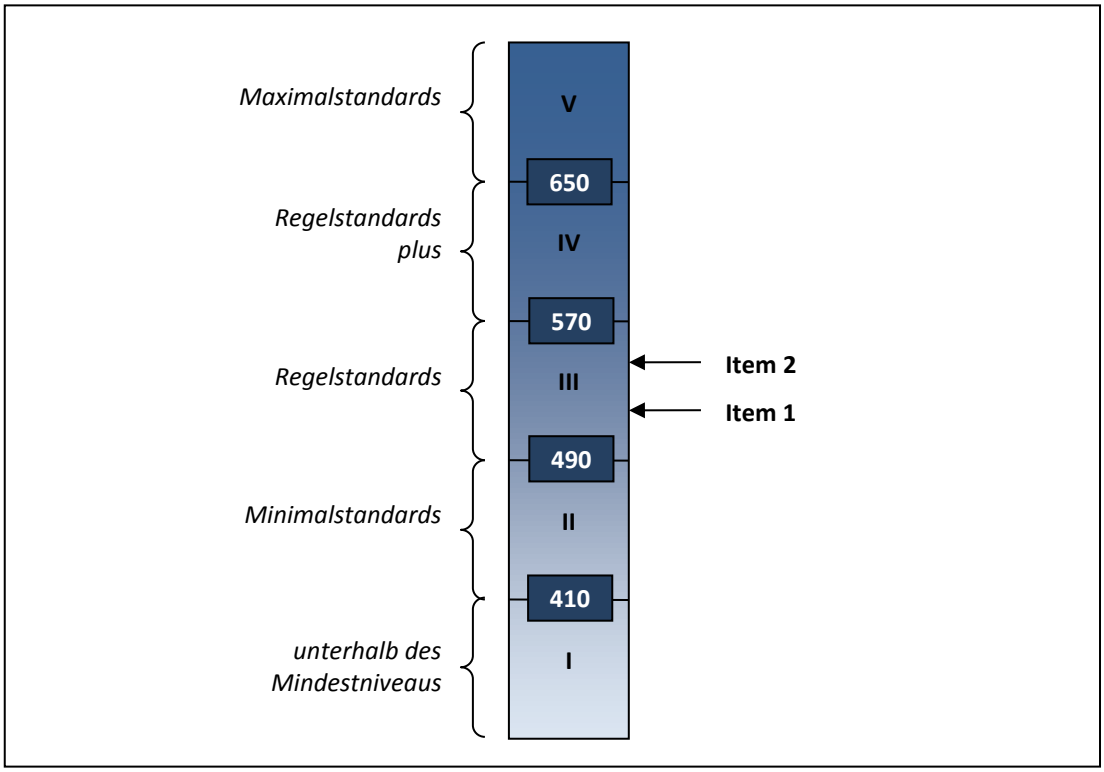

Abbildung 9:Erweiterung des Kompetenzstufenmodells durch Scores zur Zuordnung von Items (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2008, S. 19)

Die Kompetenzstufen und die Score-Skalen können auf diese Weise gemeinsam abgebildet werden. Die hier verwendete Zuordnung ist dem Kompetenzstufenmodell im Fach Mathematik für den Mittleren Schulabschluss entnommen. Im zuvor konstruierten Beispiel werden sowohl dem Item 1 mit etwa 520 Scores, als auch dem Item 2 mit circa 540 Scores die

Kompetenzstufe III (Regelstandards) zugeweisen. Diese Technik wird für jedes Item angewandt, so dass später die erreichte Kompetenzausprägung eines Schülers über seine Lösung und den Schwierigkeitsgrad des entsprechenden Items ermittelt werden kann.

Das vorgestellte dichotome Rasch-Modell wird bei Aufgaben benutzt, deren Lösungen lediglich in richtig oder falsch kodiert werden können. Bei mehr als zwei Lösungskategorien wird das Partial-Credit-Modell herangezogen, bei welchem sich für jede einzelne Kategorie eine ICC ergibt, welche in diesem Fall als Category Characteristic Curve (CCC) bezeichnet wird. Die verschiedenen CCCs eines Items können in einer gemeinsamen Grafik dargestellt werden. Die Aufgabenschwierigkeit ist durch die Schnittpunkte zwischen zwei CCCs determiniert, wo die Lösungswahrscheinlichkeit für beide Kategorien gleich groß ist (vgl. Baumert, Bos, & Lehmann, 2000, S. 64 f.).

Bei den Vergleichsarbeiten werden die Rasch-Modelle um multinomiale Rasch-Modelle erweitert. Hierbei wird die Personenfähigkeit nochmals in individuelle, die Leistung erklärende Hintergrundvariablen zerlegt, den sogenannten Plausible Values. Dies dient der Bereinigung von Messfehlern und dem „fairen Vergleich“, indem die Fähigkeit eines Schülers nicht nur unter Berücksichtigung seiner latenten Leistung und der Aufgabenschwierigkeit interpretiert wird, sondern Hintergrundindikatoren hinzugezogen werden (vgl. Abschnitt 4.3.6.2). Typische Plausible Values sind beispielsweise die Schulform, die Klassenzusammensetzung, das Geschlecht und der sozioökonomische Hintergrund der Schüler (vgl. Hartig & Kühnbach, 2006, S. 30; Hartig, 2007, S. 84).

Es existieren jedoch noch weitere Einflussfaktoren, die Messfehler verursachen können. Hierzu zählen unter anderem die Rahmenbedingungen während der Testsituation, wie Missverständnisse beim Verstehen einer Aufgabe oder unterschiedliche Hilfestellungen durch die Lehrer als Testleiter (vgl. Peek & Dobbelstein, 2006, S. 50). Diese Faktoren können bei der Skalierung und der späteren Auswertung nicht berücksichtigt werden.

Trotz der Möglichkeit der Zuordnung von Items zu Kompetenzstufen ist zu betonen, dass beispielsweise die Mathematikaufgaben zusätzlich den Anforderungsbereichen I bis III zugeordnet werden. Dies erfolgt während der Aufgabenkonstruktion auf Grundlage der Lehrerfahrung der Testkonstrukteure und wird anschließend in der Pilotierung nochmals überprüft. Das dargestellte Verfahren der Skalierung findet dabei keinen Einsatz, da ein unmittelbarer Zusammenhang zwischen den Anforderungsbereichen und den Kompetenzstufen nicht hergestellt wird.

In Rückgriff auf die Ausführungen in Abschnitt 2.3.2.2 sollte an dieser Stelle ebenfalls erwähnt werden, dass die erläuterte Vorgehensweise zur Bestimmung der Aufgabenschwierigkeit und der Zuordnung zu einem Niveau auch zur Konstruktion der Kompetenzstufen-

modelle angewandt wurde. Über die Pilotierung und Skalierung einer hinreichend großen Zahl von Items wurde deren Aufgabenschwierigkeiten ermittelt. Anschließend erfolgte die Zusammenfassung von Aufgaben mit sich ähnelnden Werten zu einer Kompetenzstufe, so dass die Niveaus prinzipiell Intervalle von Aufgabenschwierigkeitswerten darstellen (vgl. Neubrand, 2004, S. 88). Hierbei ist es unerheblich, welche kognitiven Fähigkeiten gefordert werden, sondern einzig der bei der Skalierung ermittelte Wert für die Schwierigkeit eines Items ist ausschlaggebend. Mit Hilfe dieses Vorgehens können die Grenzen zwischen Kompetenzstufen mit Scores versehen werden. Dies ermöglicht es zudem, für die Schwellen Ankerbeispiele anzugeben, welche den Übergang von einer Kompetenzstufe in die nächst höhere verdeutlichen. Ein höheres Niveau schließt somit ein niedrigeres ein und stellt darüber hinaus zusätzliche erhöhte Anforderungen (vgl. Peek & Dobbelstein, 2006, S. 47).

Je geringer die Aufgabenschwierigkeit, desto einfacher ist die Zuordnung zu einer Kompetenzstufe möglich. Der Grund hierfür ist in der erhöhten Komplexität der Items zu sehen, die mit einem steigenden Schwierigkeitsgrad einhergeht. Zur Lösung sind oftmals mehrere Bearbeitungswege unterschiedlichen Anforderungsgrades möglich, so dass es bei der Auswertung später unklar ist, welche Fähigkeiten nun gemessen werden. Folglich kann eine eindeutige Zuordnung zu einer Kompetenzstufe eventuell nicht mehr vorgenommen werden (vgl. Meyerhöfer, 2005, S. 176). Ein größeres Maß an Offenheit der möglichen Lösungswege erhöht den Entscheidungsspielraum bei der Aufgabenbewältigung, was zugleich die Aufgabenschwierigkeit beeinflusst. Dies steht wiederum in Abhängigkeit zum Komplexitätsgrad der erforderlichen Lösung (vgl. Schweitzer, 2007, S. 26).

Des Weiteren kann der Schwierigkeitsgrad einer Aufgabe durch die Fragestellung zusätzlich beeinflusst werden. Geschlossene Itemformate werden bei der Beantwortung schwerer, wenn die Falschantworten dennoch möglichst plausibel erscheinen (vgl. Baumert, Bos, & Lehmann, 2000, S. 102). Offene Items beanspruchen demgegenüber in der Regel einen noch höheren Arbeitsaufwand und Anspruchsgrad, da bei diesen nach Baumert, et al. „vorhandene Wissensstrukturen neu geordnet und in Beziehung zueinander und zur Aufgabe gesetzt werden müssen“ (2000, S. 103). Ein dritter beeinflussender Indikator ist die Konzeptschwierigkeit, welche sich zum Beispiel in der sprachlichen Gestaltung der zu bearbeitenden Texte ausdrückt. Zudem können Positionseffekte der Items in den Testheften aus der Pilotierung die zu berechnende Lösungsschwierigkeit erhöhen. Items am Ende eines Testheftes werden aus Zeitgründen von durchschnittlich weniger Schülern erfolgreich bewältigt, als wenn sie am Anfang des Testheftes positioniert worden wären (vgl. Ehmke, Leiß, Blum, & Prenzel, 2006, S. 231).

4.3.4 Durchführung, Kodierung und Ergebniseingabe

Im Anschluss an die Skalierung stellt die Entwicklergruppe die endgültigen Testhefte für den jeweils nächsten Durchlauf zusammen. Hierfür muss entschieden werden, welche fachlichen Kompetenzbereiche geprüft werden sollen. Um die einzelnen Domänen möglichst differenziert messen zu können, ist es oftmals nicht sinnvoll, alle Kompetenzbereiche in den Test zu integrieren. Des Weiteren nehmen die Testersteller eine Verteilung der Schwierigkeitsgrade für VERA 6 und VERA 8 vor. Bei diesen beiden Messungen werden drei Testhefte konstruiert: Jeweils eines für den Hauptschul-, den Realschul- und den Gymnasialbereich. Die Schüler des Realschulzweigs erhalten das sogenannte Basisheft. Im Test für die Hauptschule ist die Anzahl an leichten Aufgaben erhöht, für das Gymnasium sind dagegen quantitativ mehr schwerere Items eingebaut. Folglich wird bereits hier eine Modifizierung hinsichtlich des Anforderungsniveaus vorgenommen (vgl. Korngiebel, 2009, S. 66).

Nach der Zusammenstellung werden die Testhefte an die Schulen versandt. Die Durchführung findet an einem zentral festgelegten Tag statt. An den Vergleichsarbeiten nehmen die allgemeinbildenden Schulen sowie Grundschulen mit Förderstufen teil. Förderschulen partizipieren bisher nur, wenn sie mit den übrigen Schulformen zielgleich arbeiten und sie eine Teilnahme ausdrücklich wünschen. Reguläre Durchläufe an Förderschulen sind zwar für die kommenden Jahre geplant, das Testmaterial muss jedoch zuvor an die besonderen Schülerausgangslagen angepasst werden. Von den Vergleichsarbeiten vollkommen ausgeschlossen sind hingegen Lernhilfeschulen (vgl. ebd., S. 69). Die Entscheidung bezüglich der Teilnahmebedingungen treffen die jeweiligen Landesinstitute. Während in einigen Bundesländern die Durchführung der Vergleichsarbeiten verpflichtend ist, können in anderen Ländern die Schulen selbst über die Teilnahme bestimmen, so dass hier das Prinzip der Freiwilligkeit wirkt.

Für die Durchführung der Vergleichsarbeiten ist eine feste Bearbeitungszeit vorgesehen. Die Aufsicht führende Lehrperson erhält bezüglich des Ablaufs ein Manual mit Instruktionsanweisungen. Je präziser diese Handreichung konstruiert ist, desto stärker kann Objektivität gewährleistet und die Vergleichbarkeit erhöht werden. Für die Lehrkraft selbst ist eine detaillierte Instruktion vor allem dann wertvoll, wenn sie sich erstmalig mit dem Testinstrument konfrontiert sieht. Durch die Anweisungen zu jeder Handlung wird die Scheu vor dem Umgang mit der Leistungsmessung abgebaut. Die Sicherstellung einer objektiven Durchführung sollte bei der Erstellung solcher Hinweistexte die oberste Priorität einnehmen.

Unmittelbar nach dem Test korrigiert und kodiert der Fachlehrer die Schülerarbeiten nach einem festgelegten Bewertungssystem. Als Handreichung erhält er wiederum ein Manual,

in welchem die Kriterien für die Punktevergabe vorgegeben werden. Unabhängig vom Unterrichtsfach und den Aufgabenformaten muss die Lehrperson die folgenden Korrekturhinweise beachten (vgl. Thüringer Kultusministerium, 2009, S. 2):

- Die Schülerantworten werden entweder als richtig oder falsch klassifiziert. Mischformen oder halbe Bewertungspunkte sind nicht vorgesehen.
- Das Urteil über die Punktevergabe muss unabhängig von der Bewertung der Antworten zu anderen Items des Schülers stehen.
- Satzfragmente sind ebenfalls als richtig einzuschätzen, wenn alle geforderten Inhalte in vollem Umfang enthalten sind.
- Zeichensetzung und Rechtschreibung werden in der Regel (Ausnahmen sind gekennzeichnet) bei der Bewertung außer Acht gelassen.

Bei geschlossenen und halboffenen Items wird für jede korrekte Antwort eine Bewertungseinheit erteilt. Problematisch ist die Kodierung von offenen Items: Im Abschnitt 4.3.1.3 wurde bereits angesprochen, dass zuvor eine Signierung durch die Testentwickler vorgenommen wird, bei welcher ein Repertoire an Lösungsmöglichkeiten gesammelt und bezüglich der späteren Punktevergabe klassifiziert wird. Dieses kann als Vergleichsmaßstab für die tatsächliche Schülerleistung herangezogen werden. Für die Objektivität und Erleichterung des Bewertungsprozesses ist eine möglichst differenzierte und detaillierte Signierung von erheblicher Bedeutung. Auf diese Weise können subjektive Bewertungen jedoch nicht ausgeschlossen, sondern lediglich reduziert werden.
Nach der Korrektur gibt die Fachlehrkraft die Ergebnisse anonymisiert in einem geschützten Online-Portal ein. Vorab erhält er als Hilfestellung einen Erhebungsbogen, in welchem für jeden Schüler die erreichte Punktzahl und zusätzliche Angaben über den Lernstand, sowie vorherige Diagnosen von Förderbedarf vermerkt werden. Aufgrund des Datenschutzgebots werden die Schülernamen durch einen computergenerierten Buchstabencode ersetzt (vgl. Korngiebel, 2009, S. 71 f.).

4.3.5 Auswertung

Nachdem die Testergebnisse in das Online-Portal eingegeben wurden, folgt die Auswertung der Messergebnisse. Zum einen wird hierbei untersucht, wie die Schüler prozentual zur möglichen Gesamtpunktzahl in spezifischen Bereichen abgeschnitten haben. Darauf basierend werden Klassenanalysen erstellt und Landesmittelwerte berechnet, was im nachfolgenden Abschnitt 4.3.6 thematisiert wird. Zum anderen sollte die Verteilung der Lerngrup-

pe auf die Kompetenzstufen abgebildet werden. Ein kodierter Testwert einer Person allein hat hierbei keinen Aussagewert. Es muss eine kriteriumsorientierte Interpretation unter der Fragestellung stattfinden, was ein niedriger oder ein hoher Testwert in Bezug auf Kompetenzen und Bildungsstandards bedeutet. Das Ziel besteht somit in der Zuordnung der Testwerte einer Lerngruppe zu den spezifischen Kompetenzstufen (vgl. Baumert, Bos, & Watermann, 1999, S. 41). Die Grundlage bildet das in Abschnitt 4.3.3 beschriebene Kompetenzstufenmodell mit der zugehörigen Score-Skala, anhand welcher alle Testaufgaben einer Kompetenzstufe zugeordnet werden können. Bei der Auswertung werden die Ergebnisse derjenigen Items zusammengefasst, die sowohl auf einer gemeinsamen Kompetenzstufe als auch im gleichen Kompetenzbereich verortet sind. Mit diesen Voraussetzungen wird das Testergebnis zur erworbenen Kompetenzausprägung der Schülergruppe in Beziehung gesetzt. Bei Kenntnis der Personenfähigkeit und der Itemschwierigkeit können spezifische Vorhersagen getroffen werden, wie die Personen Aufgaben der gleichen Schwierigkeit lösen würden. Wurde eine Itemschwierigkeit bewältigt, haben die Schüler die jeweilige Kompetenzstufe mit hinreichender Sicherheit und unter Einbezug einer gewissen Fehlerquote erreicht (vgl. Bühner, 2006, S. 300; Baumert, Bos, & Lehmann, 2000, S. 116). Die ermittelten Skalenwerte können auf diese Weise inhaltlich interpretiert werden. Aufgrund der hierarchischen Abstufung der einzelnen Niveaus können folgende Schlussfolgerungen gezogen werden (vgl. Gasteiger, 2007, S. 29):

- Wenn ein Schüler die Items der Kompetenzstufe I nicht lösen kann, dann kann er auch nicht die der höheren Stufen lösen und
- wenn ein Schüler die Items der Kompetenzstufe V lösen kann, dann kann er auch die der niedrigeren Stufen lösen.

Die Lehrkräfte erhalten somit bei der späteren Rückmeldung der Ergebnisse Informationen darüber, wie sich die Leistungen der Lerngruppe auf die Kompetenzniveaus verteilen. Darauf aufbauend ist es das Ziel, Prognosen für die weitere Lernentwicklung abzugeben, so dass ersichtlich wird, inwiefern die Schüler die abschlussbezogenen Standards innerhalb der nächsten Schuljahre erreichen werden. Bisher müssen die Stufenmodelle hierfür gewissermaßen umgerechnet werden. Da für den Mittleren Schulabschluss die Regelstandards (Kompetenzstufe III) erreicht werden müssen, müssten sich die Schüler in der Klassenstufe 8 noch auf der Stufe II bzw. auf der Schwelle zur Stufe III befinden. Ein genauer Score-Wert lässt sich hierfür nicht festlegen. Das IQB schlägt diesbezüglich vor, eigene Stufenmodelle für die achte Jahrgangsstufe zu entwickeln, so dass mit den Vergleichsarbeiten der Kompe-

tenzstand der Klasse und die abzuschätzende weitere Entwicklung präziser ermittelt werden können (vgl. Institut für Qualitätsentwicklung im Bildungswesen, 2008, S. 44).
Die Analyse der Kompetenzverteilung kann bei den Vergleichsarbeiten allerdings nur klassenbezogen stattfinden. Es werden zwar auch Angaben getroffen, wie erfolgreich der Einzelschüler prozentual zum Maximalwert abgeschnitten hat, die statistische Kompetenzermittlung auf Individualebene sei nach Bos & Voss (vgl. 2008, S. 449 ff.) jedoch mit gravierenden Messfehlern behaftet. Die Gründe hierfür seien einerseits in der begrenzten Itemanzahl des Testhefts zu sehen, so dass ein Schüler nicht hinreichend seine Fähigkeiten darlegen kann. Zum anderen sei die Genauigkeit der Auswertung vom Stichprobenumfang und der Streuung der Schülerleistungen stark abhängig. Auf diese Weise entstehe ein Standardschätzfehler für die einzelne Person, welcher sich lediglich mit einer größeren Anzahl an Testitems reduzieren ließe. Hinzu kämen Faktoren wie die Auswertungsobjektivität der Lehrkräfte. Daher sei eine Verortung des individuellen Schülers auf der Kompetenzskala nur ungenau mit einer Abweichung von bis zu 100 Punkten möglich. Im Extremfall könne dies einen Leistungsunterschied von bis zu zwei Schuljahren darstellen. Um Fehlinterpretationen zu vermeiden, werden aus diesem Grund Individualanalysen bezüglich des Kompetenzerwerbs bei den Vergleichsarbeiten nicht vorgenommen. Selbst eine Interpretation auf Klassen- und Schulebene ist somit messfehlerbehaftet. Dieser Aspekt spricht dafür, dass die Vergleichsarbeiten sich nur äußerst geringfügig zur Individualdiagnostik eignen.
Zudem ist eine Darstellung von Entwicklungsverläufen nicht möglich. Die Ergebnisse stellen eine Momentaufnahme des aktuellen Leistungsstandes der Schüler dar, liefern jedoch keine Aussagen für eine Outputanalyse. Hierfür wären prozessbezogene Daten zum Kompetenzerwerb in Form einer Längsschnitterhebung erforderlich, welche Veränderungen und Entwicklungen berücksichtigen würden.

4.3.6 Rückmeldung

Nach Auswertung der Schülerergebnisse werden die Daten grafisch und didaktisch aufbereitet und den betreffenden Lehrkräften als Rückmeldung zur Verfügung gestellt. Dieses Instrument stellt eine neuartige Form des Feedbacks für die Lehrpersonen dar. Generell werden unter *Feedback* Vorgänge verstanden, bei denen eine Person Hinweise zu ihrem Verhalten oder ihren Leistungen erhält (vgl. Schneewind, 2007b, S. 13). Beispielsweise bekommen die Schüler im Schulalltag permanent Rückmeldungen sowohl in schriftlicher Form, wie durch Noten, Beurteilungen und Zeugnisse, als auch in mündlicher Form, zum Beispiel mittels eines Lobs, einer Ermahnung etc. Die Lehrkräfte nehmen hingegen wesent-

lich seltener ein bewusstes Feedback wahr. Als Indikatoren für den Erfolg der eigenen Arbeit dienen oft das Schülerverhalten, Kommentierungen von Seiten der Lernenden oder die ermittelten Noten aus den Klassenarbeiten. Weitere feedbackgebende Personen sind beispielsweise Eltern, die Schulleitung oder das Kollegium. Dabei wird die Lehrkraft oftmals ausschließlich mit informellen Rückmeldungen konfrontiert, welche sich als *soziales Feedback* charakterisieren lassen. Bei dieser Form der Rückmeldung erhalten die Lehrenden von anderen Menschen oder Personengruppen Hinweise, wie die eigenen Verhaltensweisen von den Beobachtern wahrgenommen und verstanden werden (vgl. Fengler, 2009). Das Feedback verfolgt das Ziel, die betreffende Person im Lernprozess bzw. in der Weiterentwicklung ihrer Professionalität zu unterstützen und die von ihr ausgeführten Handlungsprozesse qualitativ zu verbessern. Bedeutsam ist jedoch, dass soziales Feedback stets aus subjektiv wahrgenommenen Beobachtungen resultiert (vgl. Schneewind, 2007b, S. 14).

Die Rückmeldungen der Vergleichsarbeiten können insofern als neuartige Feedbackform der Schulpraxis charakterisiert werden, dass wissenschaftlich ausgewertete Daten von einer neutralen, externen Instanz vermittelt werden. Die subjektiven informellen Rückmeldungen, welche eine Lehrkraft oftmals lediglich zufällig erhält, werden somit durch objektive Daten ergänzt. Der Kreis der feedbackgebenden Personen wird auf diese Weise erweitert. In Anlehnung an Schneewind (vgl. ebd., S. 17) werden in dieser Arbeit unter Ergebnisrückmeldungen Daten verstanden, die im Rahmen der Vergleichsarbeiten erhoben, anschließend aufbereitet und in Form von Ergebnissen schriftlich an die jeweiligen Adressaten weitergeleitet werden.

Die Ergebnisrückmeldungen sind ausdrücklich nicht als Interventionen angelegt, sondern stellen Informationen bereit, welche für Qualitätsentwicklungsprozesse genutzt werden sollen (vgl. ebd., S. 17). Die Rückmeldungen sind somit ein wesentlicher Bestandteil im Instrumentarium der Standardisierung und Qualitätssicherung, indem sie als eine Brücke zwischen dem externen Test und den forcierten Entwicklungsprozessen im Bereich des Lehrens und Lernens fungieren (vgl. Klemm, 2000, S. 8).

Zugleich stellen die Ergebnisrückmeldungen ein Mittel dar, um die mit den Vergleichsarbeiten intendierten Ziele zu realisieren. Allerdings wird an dieser Stelle die Problematik der Vielzahl an Funktionen, welche den Vergleichsarbeiten zugeschrieben wird, erneut vehement offenbar (vgl. Abschnitt 4.2). Einerseits dienen die Rückmeldungen als Impuls bzw. als Katalysator für die Sicherung und Entwicklung der Qualität in Schule und Unterricht. Indem Lehrkräfte die Rückmeldungen als Arbeitsgrundlage für die Reflexion und Rezeption der Schülerergebnisse verwenden, können sie nach Aussage von Criblez, et al. (vgl. 2009, S. 114 f.) die Wirkung ihrer Unterrichtsarbeit in Form einer Stärken-Schwächen-Betrachtung ana-

lysieren und weiterführend Maßnahmen zur Verbesserung der Lernprozesse ergreifen. Zugleich erhalten sie durch Vergleichsnormen eine Standortbestimmung der Schülerleistungen. Das objektive Feedback erweitert die eigenen Einschätzungen zu den Lernständen der Schüler, so dass die Rückmeldungen als diagnostische Hilfestellung interpretiert werden können. Auf Schulebene bezogen können die Rückmeldungen als Kommunikationsgrundlage benutzt werden, da sie Anlass für pädagogische Diskussionen zum Lehren und Lehren in den einzelnen Fachbereichen bieten (vgl. Watermann, Stanat, Kunter, Klieme, & Baumert, 2003, S. 94 ff.).

Den Lehrpersonen zeigen die Ergebnisrückmeldungen ihre Verantwortung für die Lernprozesse verstärkt auf. Sie fungieren daher auch als eine Form der Rechenschaftslegung über die Wirksamkeit und Effizienz des Unterrichts (vgl. Tresch, 2007, S. 53). Des Weiteren können die Rückmeldungen als eine Art Gegenleistung für den Arbeitsanteil der Lehrkräfte bei der Durchführung, Korrektur und Ergebniseingabe betrachtet werden. Mit der Botschaft „Für eure geleistete Arbeit erhaltet Ihr im Gegenzug wertvolle Informationen zurück!“ erfüllen die als eine Dienstleistung interpretierten Rückmeldungen den Zweck, die Akzeptanz der Vergleichsarbeiten im Kreis der Lehrpersonen deutlich zu erhöhen (vgl. ebd., S. 52).

Aus diesen Betrachtungen lässt sich schließen, dass der primäre Adressat der Rückmeldungen die individuelle Lehrkraft ist. Die Vergleichsarbeiten sollen jedoch zugleich einen Beitrag für das Bildungsmonitoring leisten. Adressat wäre in diesem Fall nicht die Lehrperson, sondern beispielsweise die Bildungsadministration. Als Konsequenz lässt sich schlussfolgern, dass schulbezogene Rückmeldungen hier nicht anwendbar wären. Daher sind Transparenz und Zielklarheit der Rückmeldungskonzepte zentrale Kriterien für die Anpassung der Gestaltung an den jeweiligen Adressaten. Hierzu können verschiedene Aggregationsniveaus und Bezugsnormen verwendet werden (vgl. Schneewind, 2007a, S. 372). Die Leitfrage lautet: „Welche Personen erhalten welche Informationen und was soll damit geschehen?“, denn die jeweiligen Adressaten haben unterschiedliche Informationsbedürfnisse und weisen Interessen an verschiedenen Aufbereitungsformen der Daten auf (vgl. Breiter & Stauke, 2007, S. 387).

Als Akteure sind zum einen die Lehrer, Schulleitungen, Schüler sowie die Eltern zu nennen. Im Bereich des Bildungsmonitorings zählen hierzu auch Schulverwaltung, Schulaufsicht, bildungspolitische sowie wissenschaftliche Institute. Letztere benötigen zur Steuerung des Bildungssystems Zwischenberichte, welche reliable, aggregierte Daten der Gesamtpopulation beinhalten (vgl. Bähr, 2006, S. 133). Für die Akzeptanz und das Vertrauen auf Seiten der Lehrkräfte ist es bedeutsam, dass transparent kommuniziert wird, wer Zugang zu den Daten erhält. Aus diesem Grund ist es wenig sinnvoll, dass zum Beispiel die Schulaufsicht

und Schulverwaltung schul- und schülerbezogene Ergebnisse erhalten. Es sollte keinesfalls der Verdacht entstehen, dass die Daten zur Kontrolle herangezogen werden. Vielmehr sollte entsprechend den Anforderungen an konstruktives Feedback insbesondere auf eine vertrauenswürdige Konzeption der Rückmeldung Wert gelegt werden. Die Angst vor einer Weitergabe der Ergebnisse ohne Kenntnis über deren Verwendung könnte den Schul- und Unterrichtsprozessen in den Einzelschulen zuwiderlaufen und die Akzeptanz des Testinstruments schmälern (vgl. Kohler & Schrader, 2004, S. 8). Daher erfolgt die Auswertung und Rückmeldung ausschließlich anonymisiert und in der Regel in Kooperation mit einer wissenschaftlichen Einrichtung.

Die Lehrer sind im Rahmen der Vergleichsarbeiten die Erstadressaten, so dass sich die Rückmeldungskonzepte primär an den Interessen und Wünschen der Lehrer ausrichten. Für eine Standortbestimmung benötigen die Lehrkräfte Daten, welche auf Klassenebene aggregiert sind. Dies entspricht einer summativen Darstellung der Ergebnisse. Für die konkrete Unterrichtsentwicklung werden jedoch ebenso formative Elemente benötigt, die detaillierte Hinweise auf die inhaltliche Weiterarbeit und auf die Förderung spezifischer Schülergruppen liefern. Für eine konstruktive Analyse der Rückmeldung wären hierbei vor allem individualdiagnostische Informationen hilfreich (vgl. Schneewind, 2007b, S. 53).

Die Schulleitung wird hingegen die Ergebnisse aus den Vergleichsarbeiten vorrangig zur Standortbestimmung und Qualitätssicherung verwenden. Aggregationen der Daten auf Klassenebene entsprechen insbesondere dieser Anforderung. Allerdings ist in Hinblick auf das Kontrollempfinden bei den Lehrkräften eine offene Kommunikation darüber notwendig, welche Daten die Schulleitungen zu welchem Zweck erhalten.

Es bleibt zu thematisieren, inwiefern die Schüler und deren Eltern eine Rückmeldung erhalten sollten. Diese beiden Akteure finden in den derzeitigen Rückmeldekonzepten der einzelnen Bundesländer bislang nur sehr geringe Beachtung, so dass sie auf eine Übermittlung der Ergebnisse durch die Lehrkräfte angewiesen sind (vgl. ebd., S. 53). Dies ist insbesondere erstaunlich, da es die Schüler sind, die getestet und deren Leistungen beurteilt werden. Zugleich sollen die Ergebnisse dazu verwendet werden, die Lernprozesse dieser Personengruppe zu sichern und weiterzuentwickeln. Aus diesem Grund erscheint es als offensichtlich, dass die Schüler und ihre Eltern ein Interesse an individualisiert aufbereiteten Ergebnissen aufzeigen. An dieser Stelle kann durchaus eine Verantwortungsdiffusion entstehen: Sollten die Lehrer die Ergebnisse primär zur Rechenschaft ihrer geleisteten Arbeit nach außen verwenden, für sich selbst als ein Feedback zu ihrer unterrichtlichen Professionalität wahrnehmen oder die einzelnen Schüler in ihren Lernprozessen unterstützen?

4.3.6.1 Anforderungen an Rückmeldesysteme

Die Rückmeldungen dienen als Impulsgeber für die Einleitung von Maßnahmen zur Optimierung von Lernprozessen. Dennoch ist zu beachten, dass sie lediglich Informationen darlegen, die keine Interpretationen, Begründungen oder Wirkungszusammenhänge der Schülerleistungen zu der Unterrichtsqualität beinhalten. Die Ergebnisrückmeldungen können somit erst ihren Nutzen entfalten, wenn sie reflektierend betrachtet und analysiert werden. Das Feedback allein führt keineswegs zu Veränderungsprozessen. Vielmehr verfolgt es die Aufgabe, die Lehrkraft zu kritischen Betrachtungen ihrer eigenen Arbeit anzuregen. Um dies leisten zu können, müssen die Rückmeldungen spezifischen Anforderungen genügen.
Mit der verstärkten Einführung von standardisierten Testformaten entwickelte sich im deutschsprachigen Raum zugleich ein großes forschungswissenschaftliches Interesse an der Wirkungskraft von Ergebnisrückmeldungen. Die systematische Evaluation solcher Konzepte befindet sich noch im Anfang, jedoch wurden bereits über wissenschaftliche Studien sowie über die praktisch gesammelten Erfahrungen der Landesinstitute Qualitätskriterien erarbeitet, die zu einer Verwendbarkeit und Akzeptanz der Rückmeldungskonzepte beitragen (vgl. Kohler & Schrader, 2004, S. 6 f.). An dieser Stelle werden insbesondere die Zusammenstellungen von Criblez, et al. (vgl. 2009, S. 114 ff.), Posch (vgl. 2009, S. 124), Schneewind (2007b, S. 68 ff.), Tresch (vgl. 2007, S. 334) aufgegriffen, welche große Übereinstimmungen aufweisen. Daher erfolgt nun eine zusammenfassende und teilweise ergänzende Anführung der Qualitätskriterien. Die Kriterien werden abschließend in der Abbildung 10 nochmals grafisch dargestellt.

- *Vertraulichkeit und Anonymisierung*
  Wie bereits erwähnt wurde, ist es für die Akzeptanz der Rückmeldungen im Kreis der Lehrkräfte enorm wichtig, dass sie von einer vertrauenswürdigen Quelle stammen und die Daten anonymisiert ausgewertet und rückgemeldet werden.

- *Klare Kommunikation von Zielen und Grenzen*
  Den Anwendern muss transparent dargelegt werden, welche Ziele mit der Rückmeldung verfolgt werden. Zugleich müssen sie über die Grenzen in der Aussagekraft der Informationen aufgeklärt werden. Dies beugt zum einen unrealistischen Erwartungshaltungen des Anwenders vor, zum anderen werden hierdurch Missverständnisse und eine missbräuchliche Nutzung reduziert (vgl. Hosenfeld & Groß Ophoff, 2007, S. 362).

- *Einfache Übermittlung*
  Im Sinne des Verständnisses der Rückmeldung als eine Dienstleistung für die Akteure in der Schule ist es für deren Akzeptanz erforderlich, dass die zur Verfügung gestellten

Informationen leicht zu erhalten sind. Dies geschieht bei den Vergleichsarbeiten über das Online-Portal, mit dem die Lehrpersonen bereits aufgrund der Ergebniseingabe vertraut sind.

- *Zeitnahe Rückmeldung*

  Um eine Anschlussfähigkeit der getesteten Inhalte an die Unterrichtsprozesse zu gewährleisten, ist der zeitliche Abstand zwischen dem Durchführungszeitpunkt der Vergleichsarbeit und der Ergebnisrückmeldung so gering wie möglich zu halten.

- *Verständlichkeit*

  Für eine intensive Nutzung der Rückmeldungen ist eine benutzerfreundliche Gestaltung erforderlich. Dies betrifft zum einen formale Aspekte wie den Umfang, die Übersichtlichkeit oder die grafische Darstellung von Informationen. Zum anderen müssen die Informationen verständlich formuliert sein. Dies ist umso bedeutsamer, da es sich um statistische Daten handelt und hierzu kein Grundwissen von Seiten der Lehrkräfte vorausgesetzt werden kann. Nach Aussage von Peek, et al. (vgl. 2006, S. 229) muss die Komplexität der Informationen angemessen reduziert werden, indem beispielsweise einfach zu kommunizierende Kennwerte verwendet werden bzw. das statistische Grundverständnis in begleitenden Materialien vermittelt wird. Eine Überforderung soll vermieden und die Verwendbarkeit der Rückmeldungen ohne besondere Einarbeitung gewährleistet werden.

- *Zielgruppenorientierung*

  Für die Konstruktion verwendbarer Rückmeldungen ist es elementar, dass die Rückmeldungen an die jeweiligen Adressaten inhaltlich sowie in ihrer Gestaltung angepasst werden.

- *Relevanz*

  Aus der adressatengerechten Gestaltung leitet sich ab, dass die Rückmeldungen Informationen enthalten sollten, welche dem Benutzer tatsächlich nützen. Die Daten werden nur dann für einen reflexiven Lernprozess verwendet, wenn der Anwender die ihm zur Verfügung gestellten Informationen für sich persönlich als bedeutsam einschätzt.

- *Bezugsnormen und fairer Vergleich*

  Für eine Verortung und Beurteilung der Schülerleistungen müssen die Daten zu spezifischen Bezugsnormen in Beziehung gesetzt werden. Hinsichtlich eines sozialen Vergleichs sollten bei der Auswertung Hintergrundinformationen berücksichtigt werden, auf die Schule und Unterricht einen geringen bis keinen Einfluss haben. Dieser soge-

nannte „faire Vergleich“ sowie die möglichen Bezugsnormen werden nachfolgend im Abschnitt 4.3.6.2 erläutert.

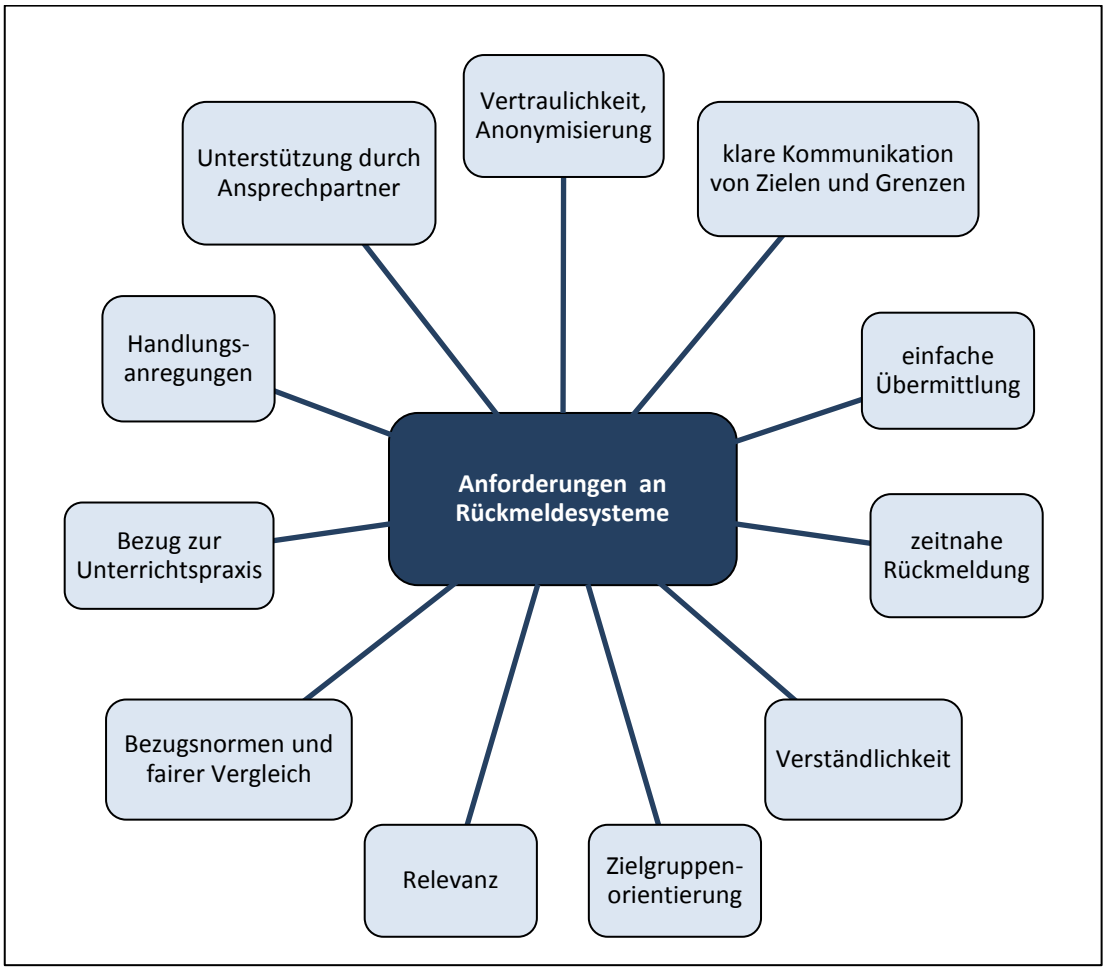

Abbildung 10: Anforderungen an Rückmeldesysteme

- *Bezug zur Unterrichtspraxis*

  Für eine Verbesserung von Prozessen muss sich Feedback stets auf beeinflussbare und veränderbare Aspekte des unterrichtlichen Agierens beziehen. Daher sollten Rückmeldungen für Lehrkräfte auf deren Handlungsfeld unmittelbar übertragbar und anschlussfähig sein.

- *Handlungsanregungen*

  Um den Reflexionsprozess zu unterstützen, muss ein konstruktives Feedback Handlungsanregungen in Form von Hinweisen enthalten. Die Rückmeldung sollte einen motivierenden und unterstützenden Aufforderungscharakter besitzen.

- *Unterstützung durch Ansprechpartner*

  Für die Akzeptanz des Gesamtkonzepts ist es enorm wichtig, dass die Lehrkräfte sich nicht mit den Rückmeldungen überfordert oder allein gelassen fühlen. Es ist daher rat-

sam, konkrete Ansprechpartner für verschiedene Kontexte (organisatorisch, inhaltlich, fachlich, statistisch) auszuweisen.

Aus den Qualitätskriterien für die Ergebnisrückmeldungen lässt sich schlussfolgern, dass die Konzepte sowohl quantitative als auch qualitative Elemente beinhalten sollten (vgl. Büchter & Leuders, 2005, S. 18). Der quantitative Bereich der Rückmeldung enthält beispielsweise die grafische Aufbereitung der Daten und deren Vergleich zu Referenzwerten in Form von Tabellen und Diagrammen. Hierzu zählt ebenso die Fähigkeitsverteilung der Schülergruppe anhand der Kompetenzstufen. Zum anderen können die Ergebnisse auf Aufgabenebene aufbereitet werden, so dass Lösungs- und Fehlerhäufigkeiten, differenziert nach den einzelnen getesteten Inhaltsbereichen, analysiert werden können (vgl. Groß Ophoff, Koch, Hosenfeld, & Helmke, 2006, S. 28). Insbesondere die Interpretation der klassenspezifischen Fehlermuster kann wertvolle Anhaltspunkte für Maßnahmen zur Verbesserung der Schülerleistungen bieten.

Die Adressaten benötigen für die Auswertung jedoch eine Anleitung, wie gehaltvolle Informationen aus der grafischen Aufbereitung gewonnen und in einem späteren Reflexionsprozess zur Qualitätsentwicklung genutzt werden können. Beispiele für solche Interpretationshilfen sind fachdidaktische Aufgabenkommentierungen, Zuordnungen der Testinhalte zu den Bildungsstandards und Kompetenzmodellen, Aufzeigen typischer Fehlerquellen und Hinweise für die weitere unterrichtliche Arbeit (vgl. Büchter & Leuders, 2005, S. 18). Ergebnisberichte für die Bildungsverwaltung, -politik und -wissenschaft sollten zudem qualitative Ausführungen zu der angewandten Testtheorie und der pädagogischen sowie fachdidaktischen Konzeption der Vergleichsarbeiten enthalten, um die Aussagekraft der Tests tiefgreifend ermitteln zu können (vgl. Nachtigall & Jantowski, 2007, S. 403).

4.3.6.2 Bezugsnormen und fairer Vergleich

Für eine fundierte Standortbestimmung ist es notwendig, die ausgewerteten Schülerdaten aus den Vergleichsarbeiten mit verschiedenen Vergleichsmaßstäben in Beziehung zu setzen. Dabei wird zwischen der *sozialen*, *kriterialen*, *ipsativen* sowie *intraindividuellen Bezugsnorm* differenziert.

Soziale Bezugsnorm

Bei der sozialen Bezugsnorm werden die erbrachten Schülerleistungen mit einer normierten Vergleichsgruppe dahingehend verglichen, wo sich die beurteilte Lerngruppe innerhalb

der Gesamtverteilung befindet (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 124). Die Stärken und Schwächen der Klasse werden über die Ergebnisse der Vergleichsgruppe ermittelt, so dass sich der Lernstand an den Leistungen anderer Schüler orientiert (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 17). Der soziale Vergleich ähnelt der im Unterricht traditionell verwendeten Orientierung am Klassendurchschnitt. Das klasseninterne Bezugssystem, welches hinsichtlich der diagnostischen Informationen eine begrenzte Reichweite aufweist, wird im Rahmen der Vergleichsarbeiten somit um eine größere Referenzgruppe erweitert (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 124).

Für die qualitative Aussagekraft des sozialen Vergleichs ist die verwendete statistische Methodik von elementarer Bedeutung. Ohne eine Differenzierung der Einflussfaktoren auf die Schülerleistung ergeben, vereinfacht dargestellt, Vergleiche zwischen Daten einer Hauptschule und einem Gymnasium wenig brauchbare Informationen. Aus diesem Grund wird ein so genannter „fairer Vergleich“ im Rahmen der Plausible Value-Auswertung (vgl. Abschnitt 4.3.6.2) angestrebt, welcher die Verschiedenheit der Schüler hinsichtlich bestimmter Merkmale, die nicht durch Schule oder Unterricht beeinflusst werden können, statistisch berücksichtigt. Je stärker die Kontextbedingungen für das Lernen der einzelnen Schülergruppen voneinander abweichen, desto schwieriger lassen sich die Ergebnisse als Output effektiver Lernprozesse interpretieren. „Faires Vergleichen von Schulen bzw. Klassen bedeutet, den kausalen Effekt der Beschulung bzw. des Unterrichts als Vergleichsgröße zu verwenden. […] Um faire Vergleiche zu erzielen ist es notwendig, die Störvariablen, welche den kausalen Effekt verfälschen, zu erfassen und in adäquater Weise zu berücksichtigen“ (Nachtigall & Kröhne, 2006, S. 65 f.). Die wichtigsten Einflussfaktoren sind hierbei die Schulform und das fachspezifische Vorwissen, welches jeder Schüler mitbringt. Die Ergebnisse vorangegangener Lernabschnitte lassen sich mittels einer Querschnittsuntersuchung wie der Vergleichsarbeit nicht ermitteln, so dass dieses Merkmal nicht erfasst werden kann (vgl. ebd., S. 65 f.). Eine weitere bedeutsame Einflussgröße ist der sozioökonomische Status der Eltern, da Schüler mit geringerer Ausprägung dieses Merkmals in standardisierten Tests tendenziell schlechtere Ergebnisse aufweisen. Hinzu treten weitere Personenmerkmale wie Geschlecht und Muttersprache. Die zentralen Einflussgrößen auf die Schülerleistungen sind in Abbildung 11 aufgeführt.

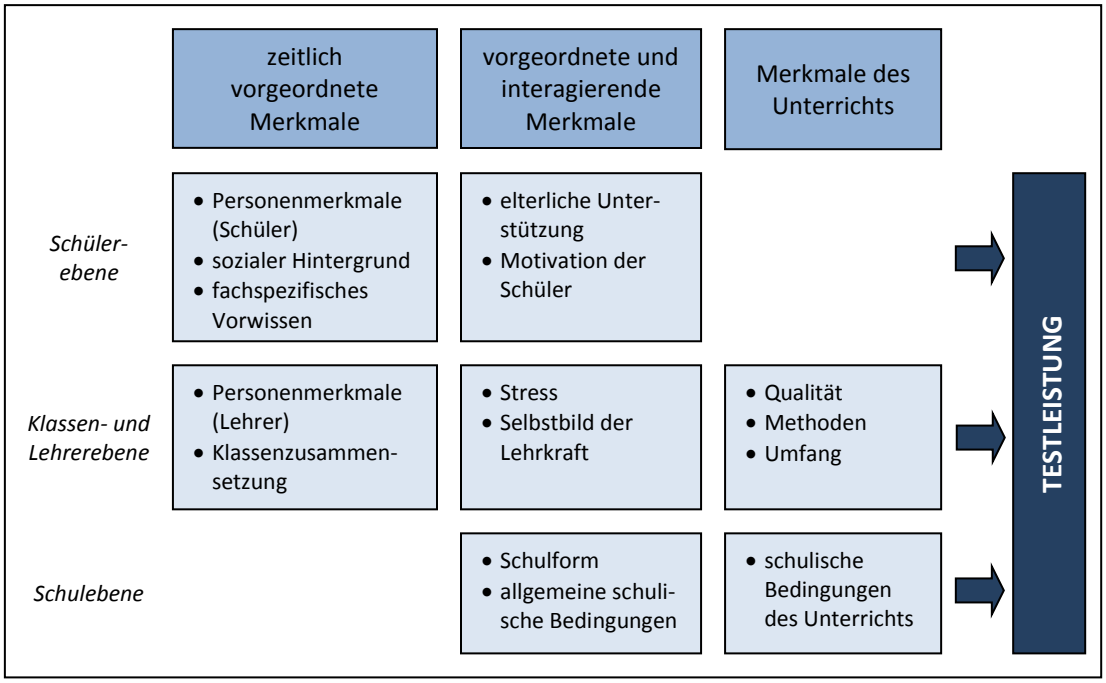

Abbildung 11: Zentrale Einflussgrößen auf Schülerleistungen (vgl. Nachtigall & Kröhne, 2006, S. 67)

Dem Schaubild ist zu entnehmen, dass nicht nur individuelle Schülermerkmale und die Schulform zu berücksichtigende Faktoren sind. Ebenso spielen die Zusammensetzung der Schülerschaft einer Klasse und die Professionalität der eingesetzten Lehrkraft eine fundamentale Rolle für den Ertrag der Lernprozesse. Dies stellt nach Nachtigall & Kröhne (vgl. 2006, S. 67) insofern ein Problem dar, als dass die Einbeziehung schulischer Merkmale in den fairen Vergleich bei der späteren Interpretation eine Unterschätzung ihres Einflusses auf die Daten zur Folge haben könnte bzw. umgekehrt die Schul- und Unterrichtseffekte überschätzt werden könnten, wenn diese Größen vernachlässigt würden. Letztlich können nicht alle Einflussgrößen im Vorfeld hinreichend identifiziert werden, so dass mit der sozialen Bezugsnorm stets nur eine Annäherung an einen fairen Vergleich möglich ist (vgl. ebd., S. 69). Die Thematik der Entwicklung von wissenschaftlich fundierten und praxistauglichen Verfahren zur Erstellung eines fairen Vergleichs bleibt daher ein relevantes Forschungsfeld.
Bei den Vergleichsarbeiten werden als Kontextvariablen Schulart, Geschlecht, Muttersprache, Anzahl der Klassenwiederholer sowie die Anzahl der Schüler mit Lernschwierigkeiten bzw. mit diagnostiziertem Förderbedarf berücksichtigt. Während diese Hintergrundmerkmale durch die betreffende Lehrperson ermittelt werden, wird die Erfassung des sozioökonomischen Status über eine an die Kinder gerichtete selbsteinschätzende Frage nach der Anzahl der Bücher im familiären Zuhause generiert. Alternativ wurden in einigen Bundesländern für jede Schulform zwei bis drei Standorttypen konstruiert, so dass sich die Schule

anhand vergleichbarer Schülerschaft und Schulumgebung selbst einem Typus zuordnen muss (vgl. Dobbelstein & Peek, 2004, S. 23 f.).

Insgesamt lassen sich vier Stufen des sozialen Vergleichs ausweisen. Zum einen können die Schülerdaten mit den einfachen Mittelwerten der Gesamtpopulation verglichen werden. Dieser „unfaire“ Vergleich ist zwar leicht zu berechnen und zu kommunizieren, qualitativ verwertbare Aussagen lassen sich jedoch nicht gewinnen. Als zweite methodische Verfahrensweise können aus allen teilnehmenden Klassen passende Vergleichsgruppen hinsichtlich der Hintergrundmerkmale herausgefiltert werden. Bei IGLU wird beispielsweise diese Methode angewandt, indem die Ergebnisse in Beziehung zu denen von vier Vergleichsklassen gesetzt werden. Drittens besteht die Zuordnung zu bereits vorher konstruierten Standorttypen. Bei der vierten Variante, welche auch bei PISA und zum Teil bei den Vergleichsarbeiten eingesetzt wird, werden nicht existierende Klassen als Vergleichsgrößen konstruiert, indem für jede einzelne Schülergruppe unter Berücksichtigung der Hintergrundfaktoren ein adjustierter Vergleichswert theoretisch ermittelt wird, wie er für die vorliegenden Kontextbedingungen zu erwarten ist (vgl. Nachtigall & Kröhne, 2006, S. 68). Obwohl dieses komplexe Verfahren in der Lehrerschaft zunächst schwieriger zu kommunizieren ist, kann hiermit ein größtmöglicher fairer Vergleich erzielt werden.

Kriteriale Bezugsnorm

Mittels der kriterialen Bezugsnorm wird die erreichte Schülerleistung mit festgelegten inhaltlichen Zielen in Beziehung gesetzt. Unabhängig von den Leistungen anderer Lernender kann auf diese Weise eine Einschätzung vorgenommen werden, ob die Schülergruppe das Lernziel erreicht hat (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 125). Als inhaltliche Kriterien werden bei den Vergleichsarbeiten die Bildungsstandards mit ihren zugehörigen Kompetenzen und Kompetenzmodellen verwendet. Speziell über die Verortung der Schülerleistungen zu den Kompetenzstufen können Ergebnisse von Lernprozessen qualitativ beurteilt werden und Hinweise auf eventuellen Förderbedarf erschlossen werden (vgl. Koch, Groß Ophoff, Hosenfeld, & Helmke, 2006, S. 189). Daher können die Bildungsstandards vor allem über die kriteriale Bezugsnorm überprüft werden. Zum anderen werden mit diesem Vergleichsmaßstab konkrete Optimierungsoptionen für den Unterricht aufgezeigt, so dass die Möglichkeit der Verzahnung von Feedback und Qualitätsentwicklung angeregt wird.

Ipsative Bezugsnorm

Bei der ipsativen Bezugsnorm wird die Testleistung einer Person oder Gruppe mit den Vorkenntnissen bzw. mit dem Anfangszustand des Lernprozesses verglichen (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 125). Folglich wird der Entwicklungsverlauf von Lernständen ermittelt. Es sind hierzu jedoch mindestens zwei Messzeitpunkte erforderlich, so dass auf Klassenebene eine Langzeitstudie notwendig wäre. Die Rückmeldungen der Vergleichsarbeiten können daher diese Bezugsnorm für die untersuchte Klasse nicht darstellen. Für eine Individualdiagnostik auf Basis der Testergebnisse wäre dies aber ein wesentliches Element einer Rückmeldung, denn diese ist insbesondere dann effektiv, wenn sie auf den Lernfortschritt und an die individuellen Lernbedürfnisse ausgerichtet ist (vgl. Crooks, 1988).

Bei Berücksichtigung der Grenzen solcher Aussagen können ipsative Bezugsnormen bei Querschnittsdesigns wie den Vergleichsarbeiten zumindest auf Schulebene angewandt werden. Es werden zwar stets Klassen mit verschiedener Zusammensetzung und unterschiedlichen Lehrpersonen betrachtet, doch kann mithilfe des fairen Vergleichs eine Schulleitung über den Verlauf mehrerer Jahre, in welchen die Vergleichsarbeiten durchgeführt werden, eine Art Trend der Schülerleistungen feststellen. Mehr Informationen können auf Schulebene mit dieser Bezugsnorm allerdings nicht gewonnen werden. Auf Systemebene stellt sich das wiederum anders dar, denn die Adressaten erhalten vollständig aggregierte Daten einer ganzen Schülerpopulation, die ohne Weiteres mit vorangegangenen Testzeitpunkten verglichen werden können.

Intraindiviudelle Bezugsnorm

Bei intraindividuellen Vergleichen werden die Leistungen der gleichen Schülergruppe in ihren unterschiedlichen Fachdisziplinen gegenübergestellt (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 18). Bezogen auf die Vergleichsarbeiten würden beispielsweise Bezüge zwischen den Ergebnissen in Deutsch und den Leistungen in Mathematik oder der ersten Fremdsprache hergestellt werden. Die Problematik hierbei liegt darin begründet, dass der anzuwendende Vergleichsmaßstab vollkommen unklar ist, so dass lediglich unreliable und wenig aussagekräftige Informationen gewonnen werden könnten. Aus diesem Grund wird bei den Rückmeldekonzepten der Vergleichsarbeiten auf intraindividuelle Bezugsnormen verzichtet.

4.3.7 Evaluation und Perspektiven für die Weiterentwicklung des Testmodells

Die Vergleichsarbeiten sind als neuartiges Instrument der Qualitätssicherung und -entwicklung im Bildungswesen zu begreifen, welche auch nach ihrer Implementierung einer fortwährenden Überprüfung und Weiterentwicklung bedürfen. Mittels einer prozessbegleitenden Evaluation sollen sowohl die Bereiche der Aufgabenentwicklung und Testauswertung sowie die Rückmeldungskonzepte ständig verbessert und an die Informationsbedürfnisse der schulischen Akteure angepasst werden. Langfristig soll die Nutzbarkeit der Vergleichsarbeiten als Bestandsaufnahme und zugleich deren Potenzial für die Schulentwicklung erhöht werden (vgl. Helmke A. , 2007, S. 225).

In einigen Bundesländern wurde für diesen Zweck im Online-Portal die Möglichkeit eines Feedbacks eröffnet, so dass die Lehrkräfte und Schulleitungen sowohl positive als auch kritische Aspekte äußern können. Die Aufgabenentwicklungsteams erhalten hingegen mittels der Analyse der Messdaten eine Rückmeldung, inwiefern jedes Item gelöst werden konnte. Deckt sich das Resultat nicht mit vorherigen Einschätzungen, müssen Gründe hierfür gesucht und spezifische Prozesse, wie die Skalierung, einer Weiterentwicklung unterzogen werden (vgl. Korngiebel, 2009, S. 87).

Neben der notwendigen Optimierung inhaltlicher Testbestandteile und des Rückmeldesystems wird zukünftig eine mögliche Reformierung der Testdurchführung zu thematisieren sein. Insbesondere den Aufwand, welcher mit der Korrektur und Ergebniseingabe für die Lehrkräfte verbunden ist, gilt es entscheidend zu reduzieren, wenn das Testinstrument von den schulischen Akteuren akzeptiert und genutzt werden soll. Daher werden im Folgenden die Möglichkeiten computerbasierter Testdurchführung diskutiert, welche hypothetisch im Rahmen der Vergleichsarbeiten langfristig betrachtet zur Anwendung kommen könnten.

Die Arbeitsökonomie könnte beispielsweise erheblich verbessert werden, wenn die Korrekturen bzw. bereits die Lösungshefte von Computerscannern eingelesen und anschließend automatisch oder von Experten ausgewertet würden (vgl. Frey & Ehmke, 2007, S. 171). Auf diese Weise kann die aufwendige und fehlerbehaftete Korrektur durch die Lehrperson erheblich minimiert werden. Jedoch verschwindet hiermit der pädagogische Spielraum bei der Beurteilung, welchen die Lehrkräfte besitzen sollten, um die Ergebnisse angemessen einschätzen zu können. Da offene Items nicht automatisch ausgewertet werden können, besteht die Gefahr einer Reduktion dieses Aufgabenformats zugunsten geschlossener Items.

Des Weiteren wird eventuell zukünftig die Möglichkeit bestehen, die Vergleichsarbeiten direkt an Computern von den Schülern bearbeiten zu lassen. Neben Verbesserungen in der Auswertungsobjektivität und der Reliabilität aufgrund geringerer Mess- und Auswertungs-

fehler bietet eine computerbasierte Anwendung den Vorteil, neue Kompetenzbereiche erfassen zu können (vgl. Jurecka & Hartig, 2007, S. 44 f.). Mittels Simulationen, Ton- und Videosequenzen, sowie dem Einsatz eines Mikrofons könnte insbesondere die produktive Sprachkompetenz Testgegenstand werden. Da der Computer in der Lebenswelt der Schüler ein zentrales Kommunikations- und Informationsinstrument darstellt, sind zudem motivationale Auswirkungen bei der Testbearbeitung zu erwarten. Es müssten aber auch negative Effekte auf die Testleistung berücksichtigt werden, denn die Kompetenz der Computernutzung ist bei den Schülern nicht gleich stark ausgeprägt. Als nachteilig ist der finanzielle Aspekt eines computerbasierten Tests zu nennen, da sowohl die Entwicklung der Programme als auch die notwendige Ausstattung in den Schulen als besonders kostenintensiv einzuschätzen ist.

Als innovative Weiterentwicklung wird von Frey (2008), Frey & Ehmke (2007) und Jurecka & Hartig (2007) die Form des adaptiven Testens in Aussicht gestellt. „Unter *adaptivem Testen* versteht man ein spezielles Vorgehen bei der Messung individueller Ausprägungen von Personenmerkmalen, bei dem sich die Auswahl der zur Bearbeitung vorgelegten Items am Antwortverhalten des untersuchten Probanden orientiert“ (Frey, 2008, S. 262). Der Test verläuft individualisiert, indem die zu testende Person solche Aufgaben bearbeitet, welche möglichst viele diagnostische Informationen liefern (vgl. Frey & Ehmke, 2007, S. 172). Die Abbildung 12 visualisiert den Ablauf eines adaptiven Tests.

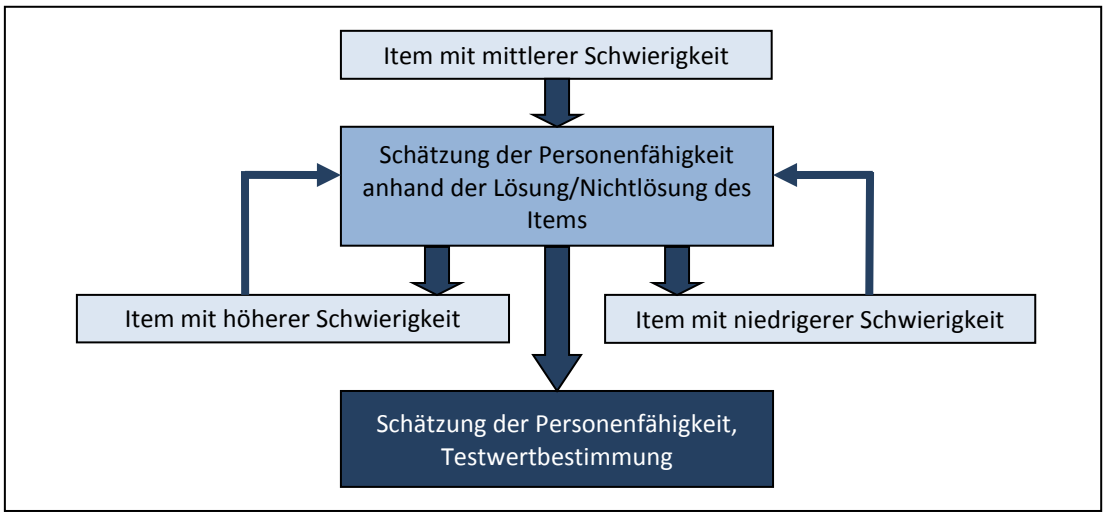

Abbildung 12: Ablaufschema eines adaptiven Tests (vgl. Jude & Wirth, 2007, S. 55)

Der Testvorgang startet mit Items von mittlerer Schwierigkeit. Im Anschluss an jede Aufgabenbeantwortung werden Schätzungen der Fähigkeitsausprägung der jeweiligen Person vorgenommen. Auf dieser Basis wird wiederum ein nächstes Item ausgewählt, das der ermittelten Fähigkeit möglichst entspricht und weitere Informationen aufzeigen kann. Es er-

folgt somit eine sehr genaue Anpassung der Testitems an das Antwortverhalten. Jeder Proband bearbeitet verschiedene Items, da die Zusammenstellung der Aufgaben erst während der Testdurchführung individualisiert erfolgt. Der Test endet letztlich, wenn genügend Informationen über die Ausprägung des Merkmals vorliegen, das heißt der Standardfehler der Personenparameterschätzung hinreichend klein ist (vgl. Frey, 2008, S. 266 ff.).

Mit Hilfe des adaptiven Testens bei den Vergleichsarbeiten könnte die Itemschwierigkeit bestmöglich an die Kompetenz der Schüler angepasst werden. Untersuchungen von Frey & Ehmke (vgl. 2007, S. 173, 182) ergaben, dass generell 40 bis 60 Prozent weniger Aufgaben bei gleicher Messpräzision bearbeitet werden müssten, so dass sich letztlich die Messeffektivität entscheidend erhöhen könnte. Zudem können Fähigkeiten in sehr hohen oder sehr niedrigen Niveaus differenzierter erfasst werden. Bislang enthielten die Vergleichsarbeiten im Paper-and-Pencil-Design viele Aufgaben mittlerer Schwierigkeit, während aus Gründen der Testökonomie der Anteil an sehr schweren oder sehr leichten Items deutlich geringer war (vgl. ebd., S. 171). Indem mithilfe adaptiver Tests die Items individuell an die Probanden angepasst werden, können auch diese Fähigkeitsniveaus präzise erfasst werden. Somit werden Über- und Unterforderungen vermieden, was wiederum positive Effekte auf die Motivation zur Testbearbeitung zur Folge hat (vgl. Frey, 2008, S. 274). Adaptive Vergleichsarbeiten würden daher einen entscheidenden Fortschritt für das Gewinnen individualdiagnostischer Informationen darstellen, da die Lehrkraft umfassende Aussagen zu den Kompetenzausprägungen eines jeden Schülers sofort im Anschluss an die Testbearbeitung erhalten würde.

So verlockend diese Weiterentwicklung der Tests zunächst klingt, desto stärker müssen die Nachteile einer solchen Testdurchführung bedacht werden. Zunächst müsste den Schülern einleuchtend dargelegt werden, warum jeder Schüler verschiedene Aufgaben bekommt und der Test bei jedem Schüler eine unterschiedliche Dauer umfasst. Zudem ist der Aspekt der erhöhten Motivation insbesondere bei sehr guten Lernenden nicht zu erwarten, da sie es gewöhnt sind, Aufgaben positiv bewältigen zu können. Bei einem adaptiven Test werden sie im Sinne einer umfassenden Fähigkeitsanalyse mit einer erhöhten Zahl an Items konfrontiert werden, welche sie nicht bearbeiten können, so dass an dieser Stelle Frustration eintreten kann (vgl. ebd., S. 275). Des Weiteren ist für diese Form der Vergleichsarbeiten ein immenser Aufgabenpool notwendig, welcher im Vorfeld entwickelt werden muss und das Spektrum der möglichen Fähigkeitsabstufungen präzise abzubilden hat (vgl. Tresch, 2007, S. 97). Offene Aufgaben müssten an dieser Stelle erneut unberücksichtigt bleiben, da sie nicht automatisch ausgewertet werden können. Hinzu treten die Probleme der Finanzie-

rung und Bereitstellung von Computern, wie sie zuvor beim computerbasierten Testen bereits angeführt wurden.

4.4 Grenzen und Risiken

Einhergehend mit der Implementierung von Vergleichsarbeiten entstand ein wissenschaftlicher Diskurs über den tatsächlichen Nutzen dieser Leistungsmessung sowie über die Gefahren und Risiken, welche mit ihnen verbunden sind. Einige dieser kritischen Argumente begründen sich aus der Testkonzeption der Vergleichsarbeiten heraus. Andere legen wiederum die Erfahrungen und Erkenntnisse aus den nordamerikanischen Staaten USA und Kanada zugrunde, welche in empirischen Untersuchungen über die Effekte standardisierter Schulleistungstests ermittelt wurden. Der offene Diskurs ist von enormer Bedeutung, damit die Grenzen der Vergleichsarbeiten transparent dargelegt und somit Fehlinterpretationen und Missverständnisse vermieden werden können. Zudem sollte den eventuellen negativen Konsequenzen der Tests möglichst präventiv vorgebeugt werden. Die zentralen Argumente zu diesen Bereichen werden im Folgenden dargelegt.

4.4.1 Einschränkungen der Aussagekraft von Vergleichsarbeiten

Im Sinne eines Überprüfungsinstruments der Bildungsstandards verfolgen die Vergleichsarbeiten das Ziel, die Lernstände von Schülergruppen zu einem bestimmten Abschnitt in ihrer Schullaufbahn möglichst valide zu ermitteln. Es stellt sich jedoch die Frage, inwiefern die Tests dies aufgrund ihrer Konstruktion überhaupt ermöglichen. Die Aufgaben- und Lösungsformate sind dahingehend konzipiert, dass lediglich das Endergebnis des Lösungsprozesses für die Einordnung in Richtig und Falsch relevant ist. Insbesondere bei den geschlossenen und halboffenen Aufgabentypen wird einzig das Ergebnis sichtbar. Die Gedankengänge der Schüler, die zur Lösung führen, werden nicht ausgeführt und somit nicht erfasst. Dies führt allerdings zu dem nicht unbedeutenden Problem, dass in keiner Weise nachvollziehbar wird, welche Denkstrategien und Lösungswege angewandt wurden. Für die Erfassung des Kompetenzstandes der Schüler ist es nicht hinreichend, lediglich die Antworten auf der Aussagenoberfläche zu betrachten. Vielmehr sind Analysen der Vorgehensweise bei der Problembearbeitung sowie die Qualität der Gedankengänge für die Interpretation der Testbearbeitung gewichtige Faktoren (vgl. Bartnitzky, 2006, S. 209; van den Heuvel-Panhuizen,

2007, S. 12). Bei offenen Aufgabenformaten wird der Lösungsweg hingegen transparenter, indem oftmals Begründungen und Argumentationen verlangt werden.

Des Weiteren ist es fraglich, inwiefern die Schüler die Kompetenzen, welche der jeweiligen Aufgabe zugrunde liegen, bei richtiger Beantwortung tatsächlich beherrschen. Je komplexer das Item konzipiert ist, desto schwieriger ist es, explizit die angesprochene Kompetenz zu bestimmen. Oftmals werden mehrere Kompetenzen in einem gewissen Verhältnis zueinander eingefordert bzw. es sind mehrere Lösungswege möglich, die verschiedene Fähig- und Fertigkeiten verlangen. Da jedoch nur das Endergebnis für die Bewertung der dargebrachten Leistung bedeutsam ist, bleibt es unklar, welcher Lösungsweg benutzt wurde und welche spezifische Kompetenz in ihrer jeweiligen Ausprägung vorliegt. Lassen sich mithilfe der Tests somit tatsächlich objektive Aussagen darüber treffen, ob Schüler mit dem identischen Endprodukt zu einem Item die gleiche Kompetenz aufweisen? Da diese Problematik nach Kreitz (vgl. 2010, S. 68) nicht zu beantworten sei, stellt sich zugleich die Frage, was die Tests letztlich überhaupt messen.

Ebenso liefert die Auswertung der Vergleichsarbeiten keine Informationen bei Nicht-Beantwortung einer Aufgabe. Wenn ein Item nicht bewältigt wurde, wobei externe Umstände wie Zeitmangel etc. in diesem Zusammenhang außen vor gelassen werden, kann die Annahme getroffen werden, dass dem Schüler eine bestimmte Voraussetzung für die Lösung fehlt. Aber lässt sich hieraus unmittelbar die Aussage ableiten, dass der jeweilige Schüler bislang noch nicht über die betreffende Kompetenz verfügt? Laut Aussage von Kreitz „kann [es], muss aber nicht am Fehlen der eigentlich interessierenden […] Kompetenzen gelegen haben“ (2010, S. 67). Bei der Auswertung der Vergleichsarbeiten wird dieses Argument jedoch nicht berücksichtigt, denn einer fehlerhaften oder nicht bearbeiteten Aufgabe werden null Punkte zugewiesen.

Auch bleibt es offen, ob die dargelegte Leistung des Schülers die „wahre“ Kompetenz darstellt. Ein Test stellt eine andere Situation wie die praktische Lebenswelt dar. Mit den Vergleichsarbeiten können zwar Informationen über die Kompetenzausprägung der Schülergruppe gewonnen werden, doch ob die Lernenden in der Lage sind, die Fähig-und Fertigkeiten in realen Problemlagen in der gleichen Ausprägung erfolgreich anzuwenden, bleibt unbeantwortet. Es werden außerdem die überfachlichen Kompetenzen, wie Kommunikation, Selbst- und Sozialkompetenzen, unberücksichtigt gelassen, die in jedem Unterrichtsfach zum Tragen kommen. Für diese Fähigkeitsbereiche existieren bereits zahlreiche Erhebungsmethoden, wie Fragebögen, Feldbeobachtungen, Interviews, Soziogramme etc., welche die Dimensionen der jeweiligen Kompetenzbereiche sehr genau erfassen können (vgl. Gaupp, 2008, S. 73 ff.; Hesse, 2008, S. 49 ff.; Reinders, 2008, S. 38 ff.; Traub, 2008, S. 65 ff.).

Diese Instrumente bleiben bislang aufgrund des ökonomischen Aufwands bei der Konzipierung der Vergleichsarbeiten außen vor.

Zudem können weitere die Testleistung beeinflussende Faktoren wie Müdigkeit, Angst, Ablenkung und Missverständnisse aufgrund der Itemformulierung nicht berücksichtigt werden (vgl. Tresch, 2007, S. 68). Gravierender ist jedoch, dass bislang höchst individuelle Merkmale, wie zum Beispiel die Intelligenz und das Vorwissen, nicht erfasst werden können, so dass dem Test automatisch die Annahme zugrunde liegt, dass die Lernprozesse bei jedem Schüler mit dem gleichem Anfangsniveau starten (vgl. Boes, 2003, S. 95). Im Rahmen des fairen Vergleichs werden lediglich förderdiagnostische Hinweise und der sozioökonomischer Hintergrund der Schüler ermittelt.

Des Weiteren soll an dieser Stelle nochmals betont werden, dass die Informationen, welche aufgrund der Vergleichsarbeiten über die Lernstände der getesteten Schüler generiert werden, nur Aussagen auf Klassenebene bzw. maximal nach verschiedenen Schülergruppen differenziert zulassen. Der Test besteht aus einer zu geringen Anzahl an Items, als dass eine Analyse der vorhandenen Kompetenzen auf Individualebene möglich wäre. Die Ergebnisse wären stark mit Messfehlern behaftet, so dass deren Rückmeldung an die Lehrkräfte und Schüler mit erheblichen Risiken in Hinblick auf eine fehlerhafte Interpretation und Verwendung der Informationen verbunden wäre.

Die Vergleichsarbeiten ermitteln die Kompetenzstände der Schüler zu einem festgelegten Zeitpunkt eines Schuljahres in Form einer Outputmessung. Daher werden lediglich die produktiven Ergebnisse von Lernprozessen erfasst. Hieraus ergeben sich zwei Probleme: Zum einen entsteht unausweichlich eine inhaltliche Diskrepanz zwischen den Unterrichtsangeboten und den Testaufgaben. Nicht alles, was im Unterricht erlernt wurde, kann getestet werden und umgekehrt ist es möglich, dass die zu testenden Kompetenzen nicht Gegenstand des Unterrichts gewesen waren. Hieraus resultiert eine curriculare Invalidität der Vergleichsarbeiten, welche die Aussagekraft über die tatsächlichen Kompetenzen der Schüler abschwächt (vgl. Herzog, 2010, S. 41). Zum anderen seien nach Regenbrecht die Messungen „diagnostisch blind“ (2005, S. 18), indem sie ausschließlich auf die Ergebnisse der Lernprozesse abzielen. Die ermittelten Resultate aus den Tests werden in keinen Zusammenhang zu den vorangegangenen Lernprozessen gebracht. Es bleibt daher unklar, unter welchen Bedingungen und Voraussetzungen die Kompetenzen erworben wurden. Aus diesem Grund lassen sich weder Entwicklungsverläufe darstellen, noch Aussagen über die Qualität des Unterrichts und der Lernsituationen ableiten (vgl. Schneewind & Kuper, 2009, S. 118; van Weeren, 2007, S. 211). Dies ist problematisch, da mithilfe der Vergleichsarbeiten die Unterrichtsentwicklung und insbesondere förderdiagnostische Elemente angeregt wer-

den sollen. Doch inwiefern ist es den Lehrpersonen möglich, dies in der Unterrichtspraxis umzusetzen, wenn der Test ihnen keinen bis nur geringen Aufschluss über die Ursachen der erbrachten Schülerleistungen liefert? Wenn der Zusammenhang zwischen Prozess und Output intransparent bleibt, besteht das Risiko, dass eine Weiterentwicklung des Unterrichts in Hinblick auf die Verbesserung der Schülerleistungen nur in Form eines Trainierens der Testformate und -inhalte stattfindet (vgl. Abschnitt 4.4.2).

Bezüglich der Einschränkungen der Aussagekraft der Vergleichsarbeiten lässt sich resümierend festhalten, dass die Validität in der Kompetenzmessung nur bedingt gegeben ist. Dies wirkt sich insbesondere zum Nachteil aus, wenn die Ergebnisse beispielsweise in der Primarstufe für die Schullaufbahnempfehlung herangezogen werden. Die scheinbare Exaktheit der Tests könne nach Heymann (vgl. 2005b, S. 8) leicht dazu führen, dass die Rückmeldungen überbewertet und andere Informationsquellen zu den Lernständen der Schüler, über welche die Lehrkräfte verfügen, abgewertet werden. „Eine Evaluation und Bewertung des Erfolgs von Unterricht allein über die Ergebnisse von Vergleichsarbeiten muss deshalb zu Verzerrungen bei der Wahrnehmung von Unterrichtsqualität und bei den Entscheidungen über Schwerpunkte der Unterrichtsentwicklung führen“ (Orth, 2001, S. 219). Die Vergleichsarbeiten sollten aus diesem Grund lediglich ein Diagnoseinstrument unter mehreren darstellen und andere Formen der Leistungsbeurteilung nicht ablösen, sondern ergänzen (vgl. Gasteiger, 2007, S. 31).

4.4.2 High-Stakes-Testing und Teaching to the Test

Mit der Einführung von standardisierten Leistungstests entstand eine fachwissenschaftliche Diskussion über deren Risiken. Insbesondere wurden hierbei die Aspekte „High-Stakes-Testing“ und das damit verknüpfte „Teaching to the Test“ problematisiert, zu welchen bereits Ergebnisse aus zahlreichen Untersuchungen zu den negativen Auswirkungen von Schülerleistungsmessungen in den USA und Großbritannien vorliegen. Die deutsche Debatte beschäftigt sich insbesondere mit der Frage, wie diese Risiken bereits bei der Implementierung vermieden bzw. zumindest reduziert werden können.

High-Stakes-Testing impliziert, dass standardisierte Tests als Basis für Sanktionen verwendet werden. Dies könne entsprechend den Ausführungen von Criblez, et al. (vgl. 2009, S. 112, 163) zum einen auf Schülerebene eine Verknüpfung der Testergebnisse mit schullaufbahnrelevanten Entscheidungen wie der Notengebung, der Versetzung in die nächst höhere Klassenstufe oder der Schulformempfehlung nach der elementaren Schulzeit bedeuten. Zum anderen könnten die ermittelten Schülerleistungen dazu verwendet werden, die Ar-

beit der Lehrpersonen zu messen und zu beurteilen. Auf Landesebene ist es hingegen möglich, mithilfe der Ergebnisse aus den Tests ein Schulranking zu veröffentlichen, welches beispielsweise als Entscheidungshilfe für die Schulwahl genutzt werden kann (vgl. Böhme, 2006, S. 9). Eine polemische Abgrenzung zwischen „guten“ und „schlechten“ Schulen wirkt sich unmittelbar auf die Schulebene aus, da sich die öffentliche Wahrnehmung von der Leistung und Qualität einer Schule transformiert. Unter Berücksichtigung der zuvor dargelegten Ausführungen zu den Einschränkungen der Aussagekraft standardisierter Tests wie den Vergleichsarbeiten erhalten die gravierenden Folgen einer einzig auf den gewonnenen Daten beruhenden Beurteilung von Schülern, Lehrkräften und Schulen zusätzliches Gewicht. Criblez, et al. (vgl. 2009, S. 112) sprechen in diesem Zusammenhang von „Zweckentfremdung“ und „Missbrauch“ der Ergebnisse, da eine defizitäre Konkurrenzorientierung stattfände.

Indem sich das Abschneiden des Schülers auf für ihn bedeutsame Entscheidungen wie Notengebung, Versetzung etc. auswirkt, erhöht sich der Leistungsdruck fundamental. Die Annahme, dass dies zu einem erhöhten Lerneinsatz, zu größerer Anstrengungsbereitschaft und letztlich zu besseren Testleistungen führe, widerlegten Amrein und Berliner (2002) in einer empirischen Untersuchung in 18 amerikanischen Bundesstaaten. Demnach ließen sich motivierende Effekte durch High-Stakes-Tests nicht nachweisen, sondern eher gegenteilige Wirkungen in Form einer erhöhten Zahl an Schulabbrechern und keiner systematischen Leistungssteigerung. „Alle High-Stakes-Tests […]“, so die Aussage von Criblez, et al., „dienen nicht primär der Förderung, sondern der Zu- und Abweisung innerhalb eines selektiven Systems“ (2009, S. 164). Würde sich dies auf die Vergleichsarbeiten in Deutschland übertragen, so wären die ursprünglichen mit den Bildungsstandards verbundenen Ziele eines Bildungsminimums, eines förderdiagnostischen, kompetenzorientierten Unterrichts sowie eines Ausgleichs der sozialen Disparitäten im Schulsystem desolat.

Insbesondere auf die Lehrerkompetenz und das professionelle Selbstverständnis von Lehrpersonen können High-Stakes-Tests fatale Auswirkungen haben. Externe Beurteilungen nahmen bislang einen sehr geringen Stellenwert im Berufsalltag der Pädagogen ein. Die Beurteilung der Qualität der eigenen Arbeit beruht meist auf Selbsteinschätzung bzw. auf schulinterner, meist unsystematischer Evaluation zwischen Kollegen oder mit der Schulleitung. Standardisierte Leistungsmessungen zwingen die Lehrpersonen nun gewissermaßen, die Resultate des eigenen Unterrichts in Form der Schülerergebnisse nach außen transparent darzulegen. Aufgrund dieser ungewohnten Situation empfinden die Lehrpersonen die Tests möglicherweise als Kontrolle. Insbesondere bei einem schlechten Abschneiden der Schüler erhöht sich der Druck. Dies kann in einen Aufschaukelungsprozess münden, so dass

die Testergebnisse als alleiniger Maßstab für die Selbsteinschätzung herangezogen werden und die herkömmlichen Indikatoren zur Vergewisserung der Qualität der eigenen Arbeit an Bedeutung verlieren (vgl. Tresch, 2007, S. 65 f.). „[U]nter diesen Umständen [ergreifen Lehrpersonen] Maßnahmen […], die eigentlich im Widerspruch zu ihrer professionellen Berufsauffassung stehen […]“ (ebd., S. 66). Nach Stähling (vgl. 2005, S. 217) ist die Bereitschaft, in einer solchen Situation neue methodische Ansätze im Unterricht zu erproben, äußerst gering. High-Stakes-Testing kann somit mehrere unerwünschte Konsequenzen nach sich ziehen (vgl. Ryan & Sapp, 2005, S. 155). Bezogen auf die Vergleichsarbeiten bedeutet dies, dass die Tests keineswegs zur Implementierung der Bildungsstandards in der Unterrichtspraxis beitragen und die Lehrkräfte auf diese Weise auch nicht zu einem kompetenzorientierten Lehren animiert würden. Vielmehr würde sich der wahrgenommene Druck auf Seiten der Lehrperson auf diejenigen Schüler übertragen, die defizitäre Leistungen im Test aufgezeigt hätten.

Darüber hinaus besteht die Möglichkeit des Betrugs, um ein höheres Testergebnis zu erreichen. Hierbei werden die Punktzahlen bei der Korrektur manipuliert oder schwächere Schüler von dem Test ausgeschlossen, so dass sich die Durchschnittsleistung erhöht (vgl. Schirp, 2006a, S. 267). Des Weiteren können die Lehrpersonen während der Testdurchführung eingreifen, indem mehr Bearbeitungszeit zur Verfügung gestellt oder vermehrt Hilfestellungen angeboten werden (vgl. Stähling, 2005, S. 218). Dies wirkt sich nicht nur auf die Durchführungs- und Auswertungsobjektivität negativ aus, sondern führt insbesondere bei einem öffentlichen Ranking zu einer ungerechten Beurteilung der „ehrlichen“ Schulen und somit zu einer Verfälschung der Gesamtergebnisse (vgl. Schirp, 2006a, S. 267).

Mit zunehmendem Druck steigt das Bedürfnis, die Lernenden möglichst optimal auf die bevorstehenden Tests im Vorfeld vorzubereiten. Der Unterrichts wird im „Teaching to the Test“ nicht als wertvolle Lernzeit genutzt, sondern für die Einübung der Aufgabenformate und Testinhalte. Je stärker das Prinzip des High-Stakes-Testing ausgeprägt ist, desto größer ist die Gefahr eines Teaching to the Test in Form eines systematischen massiven Trainings, welches den Schülern zugleich eine enorme Bedeutsamkeit ihres Testabschneidens vermittelt (vgl. Maag Merki, 2010, S. 158; Ryan & Sapp, 2005, S. 151). Dies kann zum einen curriculare Auswirkungen haben. Im Zentrum des Unterrichtsgeschehens stehen die testspezifischen Inhaltsfelder des jeweiligen Faches, so dass relevante Fähigkeiten, wie die in einer standardisierten Leistungsmessung meist nicht erfassten überfachlichen und kommunikativen Kompetenzen in den Hintergrund treten. Zum anderen tritt hiermit eine generelle Dominanz der Testfächer ein, indem die nicht-getesteten Unterrichtsfächer ebenfalls zurückgedrängt werden (vgl. Maier, 2010a, S. 115). Dies wirkt sich zugleich auf die Wahrnehmung

von Schule und Unterricht sowohl bei den schulischen Akteuren als auch in der Öffentlichkeit aus, da eine Differenzierung zwischen „wichtigen“ und „unwichtigen“ Fächern propagiert und innerhalb einer Fachdisziplin eine einseitige Gewichtung entgegen der fachdidaktischen Expertise vorgenommen wird (vgl. Bartnitzky, 2006, S. 203). Es entsteht die Gefahr einer „Verengung des Schulcurriculums“ (Schirp, 2006a, S. 268), da der Test selbst als eigentlicher Lehrplan fungiert.

Ein Teaching to the Test drückt sich zum anderen auch, je nach der Intensität seiner Anwendung, in methodischen und didaktischen Veränderungen des Unterrichts aus. Eine Vorbereitung der Schüler auf den Test gelingt am besten über ein Training anhand ähnlicher Aufgabenformate (vgl. Schirp, 2006b, S. 430). Im Vordergrund stehen Lernarrangements, in denen auf möglichst direktem Wege die betreffende kognitive Kompetenz erworben werden kann (vgl. Regenbrecht, 2005, S. 19). Besonders lehrergelenkte Methoden können das Resultat einer solchen Unterrichtsplanung sein (vgl. Maier, 2010a, S. 115). Schüleraktivierende Lernsituationen oder differenzierende Förderung erlangen hierbei zunehmend einen nachrangigen Status. Dies impliziert zugleich, dass lediglich das Lernen kurzfristigen, oberflächlichen Testwissens angestrebt wird, nicht jedoch ein tiefgehender, kumulativ aufgebauter Kompetenzerwerb (vgl. Stern & Hardy, 2002, S. 167). Indem Lernaufgaben durch Testaufgaben ersetzt werden (vgl. Abschnitt 4.3.1.2), wird die Zielstellung eines kompetenzorientierten Unterrichts und somit auch die der Bildungsstandards gänzlich verfehlt und gegenteilige Effekte impliziert.

In diesem Kontext verlieren die didaktischen und pädagogischen Fähigkeiten der Lehrkraft an Bedeutung, so dass nach Schirp (vgl. 2006a, S. 268) die zentralen Leistungstests eine Entprofessionalisierung der Lehrerschaft befördern. Die Lehrenden fungieren in der Funktion eines Testvorbereiters, dessen Aufgabe es ist, seine Schüler derart zu trainieren und motivieren, so dass sie im Test möglichst gut abschneiden.

Infolge eines Teaching to the Test können sich die Schülerleistungen tatsächlich verbessern, weil der Umgang mit dem Messinstrument routiniert und der Unterricht auf die Testanforderungen ausgerichtet wurden. Die Leistungen in den nicht-getesteten Bereichen würden jedoch aufgrund der geringer werdenden Förderung kontinuierlich abnehmen (vgl. Bartnitzky, 2006, S. 209). Für die Aussagekraft der Testergebnisse lässt sich hieraus schlussfolgern, dass weniger die Kompetenzen, sondern eher Testbewältigungsstrategien erfasst würden. Wenn standardisierte Leistungsmessungen mit Konsequenzen und Sanktionen verbunden sind, steigt die Gefahr eines Teaching to the Test, was wiederum mit einer geringeren Aussagekraft in Hinblick auf die tatsächlichen Leistungen der Schüler verbunden ist. Indem die Leistungen als Entscheidungsgrundlage verwendet werden, tritt eine Verzer-

rung der Ergebnisse ein. Die aus den Leistungen der Schüler gewonnen Informationen verlieren für eine Verortung und für förderdiagnostische Zwecke an Wert. Folglich verfälschen die Rückmeldungen die Ausgangsbedingungen für eine Unterrichtsentwicklung auf Basis der Testergebnisse (vgl. Schirp, 2006b, S. 428).

Die Ausführungen zu den Effekten von High-Stakes-Testing und Teaching to the Test lassen sich selbstverständlich nicht gänzlich auf die Vergleichsarbeiten in Deutschland übertragen. Bereits bei der Konzeption und Implementierung der Tests galt es, die möglichen negativen Konsequenzen zu vermeiden bzw. zu reduzieren. Ein Ranking ist für die Vergleichsarbeiten nach Aussage der EMSE zunächst nicht vorgesehen (vgl. Netzwerk Empiriegestützte Schulentwicklung, 2006, S. 3). Bei einer Veröffentlichung der Ergebnisse müssten die schulischen Kontextbedingungen und individuellen Lernausgangslagen unberücksichtigt bleiben, so dass ein Ranking als ein „unfairer" Vergleich keine wirkliche Aussagekraft hätte und sich eher kontraproduktiv auswirken würde. „Eine entscheidende Rolle für die Akzeptanz und für den Umgang mit den Daten in der Schulpraxis spielt das Ausmaß ihrer Veröffentlichung. Wegen der noch bestehenden Vorbehalte sollten Ergebnisse aus Lernstandserhebungen auf Schul-, Klassen- und Individualebene *zunächst* nur für den dienstinternen Gebrauch und pädagogischen Bedarf verwendet werden" (ebd., S. 4). Hingewiesen sei an dieser Stelle, dass die EMSE jedoch bei Abschluss der Implementierung und bei vorhandener Akzeptanz auf Seiten der Schulakteure eine Veröffentlichung der Ergebnisse aus den Vergleichsarbeiten nicht gänzlich ausschließt. Trotz der Empfehlung durch die EMSE gibt es in Deutschland keine einheitliche Vorgehensweise bezüglich der Verwendung und Publikation der Ergebnisse.

Es stellt sich an diesem Punkt die Frage nach der Legitimation der Veröffentlichung der gewonnenen Daten aus den Vergleichsarbeiten. Diese ist aufgrund der Diffusion bei den verschiedenen Funktionen, welche die Tests zu erfüllen haben, nicht klar zu beantworten. Sollen mittels der Vergleichsarbeiten Informationen zu den Schülerleistungen auf Länderebene im Sinne eines Systemmonitorings generiert werden, ist eine Veröffentlichung der Ergebnisse nicht nur legitim, sondern in Hinblick auf Transparenz notwendig. Hierbei sollten möglichst adjustierte Daten im Sinne eines fairen Vergleichs verwendet werden, welche im Gegensatz zu reinen Durchschnittswerten mit einer größeren Aussagekraft verbunden sind. Sollen die Vergleichsarbeiten jedoch schulintern dazu genutzt werden, die Schul- und insbesondere die Unterrichtsentwicklung zu forcieren, führt eine Veröffentlichung der Schülerleistungen zu Misstrauen gegenüber den Landesinstituten und zu einer Ablehnung des Messinstruments. Die Veröffentlichung würde den eigentlichen Zielen der Vergleichsarbeiten konträr entgegenstehen und gegenteilige Effekte bewirken, indem ein Teaching to the

Test befördert würde. Eine Bestätigung dieses Widerspruchs findet sich ebenfalls in den Ausführungen der EMSE, welche wie oben zitiert eine Veröffentlichung der Ergebnisse zwar bislang ablehnt, aber betont, dass „[in] der Perspektive einer eigenverantwortlichen Schule […] jedoch die pädagogischen Leistungen einer Schule öffentlich zu verantworten [sind] […]“ (ebd., S. 4).

Doch auch bei Nicht-Veröffentlichung der Ergebnisse ist ein Teaching to the Test nicht ausgeschlossen, da der soziale Vergleich ein Bestandteil der Rückmeldung ist, in welchem die untersuchte Schülergruppe mit dem korrigierten Landesmittelwert verortet wird. Je nach Stärke des innerschulischen Drucks, welchen die einzelne Lehrkraft persönlich empfindet, kann hieraus ein systematisches Vorbereiten der Schüler auf den Test erwachsen. Dieser negative Effekt kann abgeschwächt werden, indem die Vergleichsarbeiten möglichst innovativ gestaltet werden. Das kann mittels anwendungsbezogener Aufgaben, für die ein gezieltes Trainieren im Vorfeld kaum möglich ist, einer gehaltvollen Ergebnisrückmeldung sowie über didaktische Kommentierungen für die Weiterarbeit im Unterricht geschehen. Durch qualitativ hochwertige Tests, welche direkt mit den Bildungsstandards und kompetenzorientiertem Unterricht verknüpft werden, soll die Professionalität der Lehrer erweitert und nicht gemindert werden (vgl. Maier, 2010a, S. 117).

Eine gewisse Vorbereitung der Schüler auf die Tests kann nach Aussage von Hosenfeld, et al. (vgl. 2006, S. 71 f.) sogar förderlich sein, da auf diese Weise Störfaktoren wie Angst und Unsicherheit reduziert werden. Dies können einerseits rein organisatorische Informationen über den Ablauf der Vergleichsarbeit darstellen, andererseits können aber auch die Aufgabenformate kurz im Unterricht eingeführt werden, um Vertrautheit mit dem Testverfahren herzustellen. Des Weiteren sei es hilfreich, den Schülern Testbearbeitungsstrategien zu vermitteln, welche für sie nicht nur die Vergleichsarbeit, sondern generell als Methodenkompetenz wertvoll seien. Bei konsequenter Anwendung eines kompetenzorientierten Unterrichts sei hingegen eine inhaltliche Vorbereitung auf die Tests nicht nötig, da der langfristige Erwerb der in den Bildungsstandards geforderten Kompetenzen automatisch zu der erforderten Testleistung führe.

5 Vergleichsarbeiten und Schulentwicklung

5.1 Grundlagen der Schulentwicklungstheorie

Wie bereits in Abschnitt 2.1 ausgeführt wurde, agiert die Einzelschule im Rahmen der Entwicklung zur eigenverantwortlichen Schule als eine pädagogische Handlungs- und Gestaltungseinheit. Dies impliziert, dass die direkte Steuerung nicht durch die Administration erfolgt, sondern durch die Schule selbst in Form eines nachhaltigen Arbeitsprozesses, der Schulentwicklung. Peek definiert *Schulentwicklung* als einen „lange[n], kontinuierliche[n], dynamische[n], planmäßige[n] Analyse-, Problemlöse-, Innovations- und Lernprozess […]“ (2009, S. 1348). Schulentwicklung ist in diesem Sinne die Voraussetzung für die Qualitätssicherung und -entwicklung, indem die Arbeitsschwerpunkte stets an die jeweiligen Bedingungen und Notwendigkeiten der Einzelschule angepasst werden (vgl. Buchen, 2009, S. 39). Die Arbeitsprozesse werden von den Akteuren der Schule selbst getragen, das heißt von der Schulleitung, den Lehrkräften und der Schüler- und Elternschaft, während der Bildungsadministration und weiteren Institutionen vorrangig ressourcensichernde und unterstützende Funktionen zukommen (vgl. Rolff, 2007b, S. 13). Schulentwicklung ist sowohl auf die Gegenwart bezogen, da sie den alltäglichen Unterrichts- und Erziehungsauftrag zu erfüllen hat, als auch auf zukünftige Entwicklungen ausgerichtet (vgl. Buchen, 2009, S. 38). Das Ziel der Schulentwicklung besteht somit darin, mittels regelmäßiger Analyse und gezielter Innovationen eine bewusste Weiterentwicklung der Einzelschule voranzutreiben (vgl. Peek, 2009, S. 1348).

Die Schulentwicklung besteht nicht ausschließlich aus intrinsisch motivierten Entwicklungskonzepten, wie die Konstruktion von Schul- und Methodencurricula. Auch landes- und bundesweite Reformmaßnahmen, zum Beispiel die Implementierung der Bildungsstandards, beeinflussen die Schulentwicklung maßgeblich. Des Weiteren müssen die Bestrebungen zur Schulentwicklung nicht primär auf die Verbesserung der Unterrichtsqualität abzielen, sondern ihre Wirkungen können verschiedene Bereiche des Kontextes Schule berühren. Rolff (vgl. 2007b, S. 15 f., 29 ff.) differenziert daher das Spektrum der Schulentwicklung in die drei Dimensionen Organisationsentwicklung, Personalentwicklung und Unterrichtsentwicklung, deren Grundzüge im Folgenden kurz vorgestellt werden.

Organisationsentwicklung

Bevor eine Definition der Organisationsentwicklung erfolgen kann, sollte zunächst erläutert werden, inwiefern die Schule als eine Organisation zu verstehen ist. Organisationen sind „mehrdimensionale und offene soziale Systeme“ (Böttcher, 2008, S. 188), welche durch zentrale Zielgerichtetheit, identifizierbare Grenzen und soziale Interaktion gekennzeichnet sind. Die Organisation Schule ist hierbei auf den pädagogischen Erziehungs- und Bildungsauftrag ausgerichtet, wobei der Handlungs- und Gestaltungsspielraum durch die staatliche Kontrolle und Steuerung in gewisser Weise eingeschränkt ist (vgl. Münch, 2004, S. 27). Als eine Besonderheit der Organisation Schule kann die schwach ausgeprägte Technologiesierung der pädagogischen Arbeitsprozesse angeführt werden (vgl. Rolff, 1993, S. 121 ff.). Hinzu kommen eine vertikale und horizontale Dezentralisierung, da zum einen innerhalb der Organisation eine flache Hierarchie mit lediglich zwei Ebenen, den Lehrkräften und der Schulleitung, existiert. Dies impliziert zugleich, dass der Qualifikations- und Ausbildungsgrad der Mitarbeiter eine sehr homogene Struktur aufweist. Zum anderen verfügen die Lehrer in ihrer Unterrichtstätigkeit über eine relativ hohe Autonomie, so dass eine Kontrolle und Steuerung der Prozesse innerhalb der Organisation erschwert bzw. in einigen Bereichen kaum möglich ist (vgl. Hartung-Beck, 2009, S. 56; Münch, 2004, S. 29).

Die Organisationsentwicklung versucht die pädagogischen, technischen und menschlichen Merkmale der Organisation Schule zu integrieren, indem die Selbstentwicklung der Mitarbeiter und die Selbsterneuerung der Schule als Organisation angestrebt werden (vgl. Rolff, 1986, S. 23). Münch definiert Organisationsentwicklung daher folgendermaßen:

> *Organisationsentwicklung* umfasst „die Gesamtheit der auf die Erreichung von Zwecken und Zielen gerichteten Maßnahmen, durch die einerseits ein soziales System gebildet und strukturiert wird und andererseits die Aktivitäten der zum System gehörenden Menschen, insbesondere ihre Kommunikation und Kooperation, sowie der Einsatz von Sachmitteln geregelt werden.“ (Münch, 2004, S. 27)

Die Organisationsentwicklung ist auf die Organisation Schule als Ganzes ausgerichtet ist. Sie beinhaltet in der Regel relativ komplexe Innovationen, deren Prozesse nicht linear, sondern zyklisch verlaufen und deren Wirkungen sich auf verschiedensten Ebenen entfalten können. Daher ist die Organisationsentwicklung stets langfristig und stark prozessorientiert angelegt (vgl. Klippert, 1997, S. 13; Rolff, 2007b, S. 14). Die Bedürfnisse der Schule als Institution und ihrer Mitglieder sollten zudem gleichberechtigt gewichtet werden (vgl. Rolff, 1986, S. 23).

Organisationsentwicklung stellt in diesem Sinne nicht das Kerngeschäft der Lehrkräfte dar, sondern kann als eine Grundbedingung betrachtet werden, welche die organisatorischen Rahmenbedingungen und zugehörigen Unterrichtsprozesse erst ermöglicht. Zudem befördert sie die Weiterentwicklung der pädagogischen Professionalität, so dass sie ebenfalls eine Personalentwicklung forciert (vgl. Reh, 2008, S. 165). Klippert (vgl. 1997, S. 13) sieht im Zuge der notwenigen Klärungs- und Abstimmungsprozesse jedoch die Gefahr einer erhöhten Konferenz- und Arbeitsbelastung, von deren Ausmaß viele gutwillige Lehrkräfte abgeschreckt würden, weil sie sich durch die vielschichtige Sisyphusarbeit überfordert fühlen würden. Zielgerichtetheit, Klarheit, Transparenz und Kommunikation sind aus diesem Grund zentrale Anforderungen an die Organisationsentwicklung.

Personalentwicklung

Eine weitere Komponente der Schulentwicklung ist die Personalentwicklung. Nach Mentzel definiert sie sich wie folgt:

> „[*Personalentwicklung* ist der] Inbegriff aller Maßnahmen, die der individuellen beruflichen Entwicklung der Mitarbeiter dienen und ihnen unter Beachtung ihrer persönlichen Interessen die zur optimalen Wahrnehmung ihrer jetzigen und künftigen Aufgaben erforderlichen Qualifikationen vermitteln.“ (Mentzel, 1997, S. 15)

Mit Personalentwicklung werden somit zwei Ziele verfolgt: einerseits soll die professionelle Kompetenz der Mitarbeiter im Rahmen von Personalfortbildung, -führung und Persönlichkeitsentwicklung gefördert sowie die Arbeitszufriedenheit der Mitarbeiter gesichert bzw. gesteigert werden (vgl. Niedermair & Bachmann, 2002, S. 103). Maßnahmen der Personalentwicklung sind aus diesem Grund oftmals mit Beratung und/oder Beurteilung verbunden (vgl. Buhren & Rolff, 2000, S. 262). Andererseits soll die Personalentwicklung zugleich auch den Erfolg der Organisation Schule sicherstellen. Strategien zur Entwicklung der Professionalität der Mitglieder müssen daher den Grundsätzen von Effektivität und Effizienz folgen (vgl. Niedermair & Bachmann, 2002, S. 102 f.). Die Personalentwicklung kann durchaus individualisiert erfolgen, darf jedoch den Zielen der Organisationsentwicklung nicht entgegenstehen. Die Voraussetzung für eine wirksame Personalentwicklung ist die Bereitstellung der hierfür notwendigen Förderungsangebote.

Die Personalentwicklung erhält im schulischen Kontext eine besondere Bedeutung, da in dieser Organisationsform die sozialen Interaktionen aufgrund der geringen Technologiesie-

rung stark personengebunden sind (vgl. Terhart, 2010, S. 255). Folglich werden die Lernprozesse und -ergebnisse in einem hohen Ausmaß von den unterrichtenden Lehrpersonen beeinflusst. Meetz bekräftigt diesen Sachverhalt, indem er betont, dass„[d]ie Qualität von Bildungs- und Erziehungsprozessen und somit des Unterrichts […] im besonderen Maße von Qualifikationen und der Motivation der Lehrkräfte sowie dem Zusammenwirken der Menschen untereinander [abhängen]“ (2007, S. 122). Der Personalentwicklung sollte demnach maximale Bedeutung zukommen. Jedoch ist nach Terhart (vgl. 2010, S. 256) zu beachten, dass Personalentwicklung nur dann erfolgreich verlaufen könne, wenn die zugehörigen Maßnahmen im weitesten Sinne nicht durch die Bildungsadministration fremdbestimmt erfolgen oder als solche wahrgenommen würden. Vielmehr sollten die Mitglieder der Einzelschule bei der Entwicklung der Förderkonzepte partizipieren können, was wiederum deren Akzeptanz erhöhen kann.

Unterrichtsentwicklung

Wie bereits angeführt, sind alle Maßnahmen der Schulentwicklung an dem obersten Ziel der Erfüllung des Erziehungs- und Bildungsauftrages direkt oder indirekt ausgerichtet. Während Organisations- und Personalentwicklung nur indirekte Effekte auf dieser Zielsetzung bewirken können, setzt die Unterrichtsentwicklung unmittelbar am Erziehungs- und Bildungsauftrag an. Die Lernprozesse der Schüler vollziehen sich in der Regel in Unterrichtsarrangements, so dass der Unterricht das Zentrum der Schulaktivitäten darstellt. Aus diesem Grund ist die Unterrichtsentwicklung eine Dimension der Schulentwicklung von hoher Priorität und kann nach Horster & Rolff folgendermaßen definiert werden:

> „[Unter *Unterrichtsentwicklung* ist] die Gesamtheit der systematischen Anstrengungen [zu] verstehen, die darauf gerichtet sind, die Unterrichtspraxis […] zu optimieren.“ (Horster & Rolff, 2006, S. 60)

Das Ziel von Unterrichtsentwicklung besteht somit in der Effektivierung des Lernens in allen Bereichen (vgl. ebd., S. 60). Sie orientiert sich stets an funktionalen Anforderungen und sollte daher systematisch und fokussiert betrieben werden (vgl. Rolff, 2007b, S. 20, 132). Unterrichtsentwicklung lässt sich nicht einzig auf die Einführung neuartiger Methoden reduzieren, sondern Vernetzung, Teamarbeit, Reflexion und Evaluation sind ebenfalls bedeutsame Komponenten nachhaltiger Unterrichtsentwicklung. Insbesondere im Rahmen eines kompetenzorientierten Lernens erhält die Zusammenarbeit des Kollegiums eine neue Relevanz, wenn beispielsweise Vereinbarungen über die Vermittlung von Lernstrategien in

einzelnen Fächern bzw. Klassenstufen, die Planung von fächerübergreifendem Unterricht oder die Erstellung des Schulcurriculums erforderlich werden (vgl. ebd., S. 133, 151 ff.).
Der Unterricht selbst ist in der Gesamtorganisation der Schule jedoch der Bereich, der am geringsten durch die Schulleitung oder durch die Bildungsadministration direkt zu steuern und zu kontrollieren ist. Die Unterrichtstätigkeit ist daher für den Lehrer ein bislang relativ eigenverantwortliches Tätigkeitsfeld, welches nur durch wenige Rahmenrichtlinien, wie die Kerncurricula und die Bildungsstandards, begrenzt wird. Daher sollten die Strategien der Unterrichtsentwicklung möglichst vom Inneren der Organisation heraus entwickelt werden.

5.2 Drei-Wege-Modell der Schulentwicklung

Nach Rolff (vgl. 2007b, S. 30) besteht die Schulentwicklung aus der Synthese von Organisations-, Personal- und Unterrichtsentwicklung, dem sogenannten Drei-Wege-Modell (vgl. Abbildung 13):

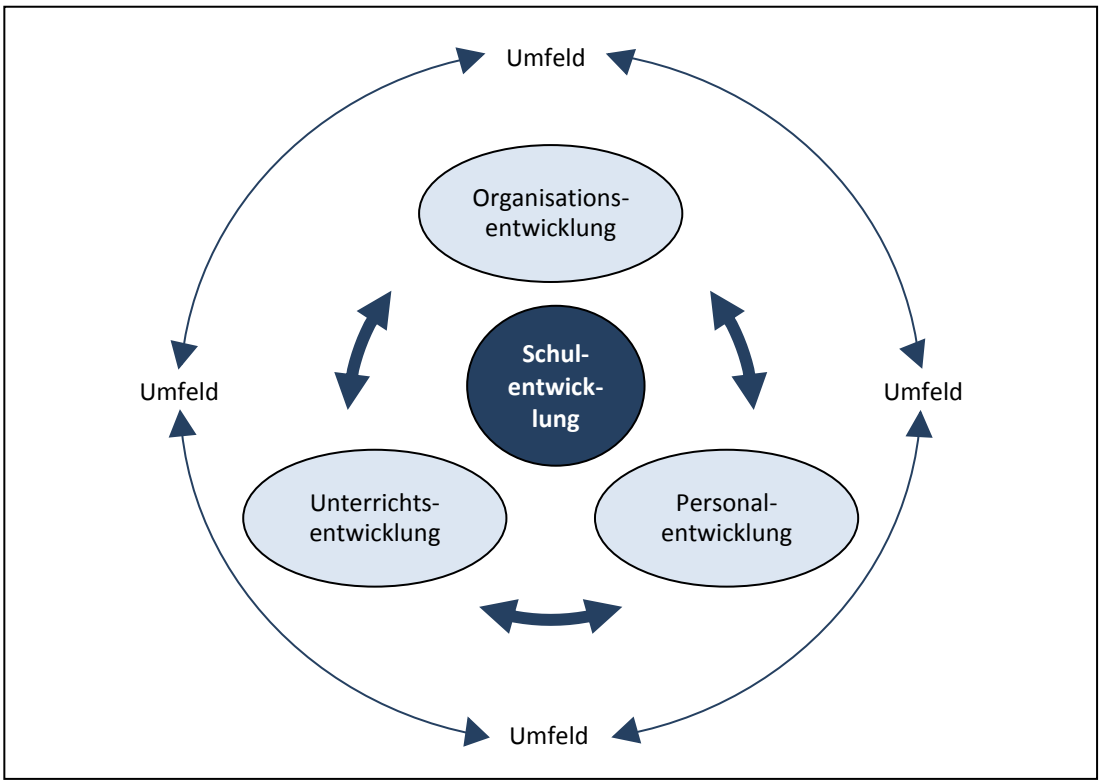

Abbildung 13: Drei-Wege-Modell der Schulentwicklung (vgl. Rolff, 2007b, S. 30)

Aus dem Modell wird ersichtlich, dass aufgrund des Zusammenwirkens von Organisations-, Personal- und Unterrichtsentwicklung ein innerschulischer geschlossener Systemzusammenhang entsteht. Dieser muss allerdings mit einer Öffnung der Einzelschule nach außen, wie zu Schulaufsicht, Schulamt, Schulträger, Gemeinde, Wirtschaft und Universitäten, ergänzt werden. Dieses äußere Umfeld wirkt in direkter Form auf die Organisations- und Per-

sonalentwicklung ein, während die Unterrichtsentwicklung aus den zuvor genannten Gründen ein relativ geschlossenes Areal bleibt (vgl. ebd., S. 31).

Dennoch bedingen sich alle drei Komponenten der Schulentwicklung gegenseitig, so dass die Förderung eines Bereichs zugleich die Entwicklung der anderen Bereiche nach sich zieht (vgl. Horster & Rolff, 2006, S. 58 f.). Beispielsweise bedarf die nachhaltige Unterrichtsentwicklung zugleich eines organisationales Lernens, indem die Verständigung unter den Lehrkräften und kollektiv geteilte Einstellungen über die Realisierung von qualitativ wertvollem Unterricht erreicht werden. Zugleich kann die Unterrichtsentwicklung nicht auf einen individuellen Kompetenzzuwachs der jeweiligen Lehrperson verzichten, welche vor der Aufgabe steht, die kollektiv geteilten Strategien in ihrem individuellen Unterricht effektiv umzusetzen (vgl. ebd., S. 60). Es können sich keine komplexen Innovationen im Kontext von Organisationsentwicklung etablieren, wenn damit nicht zugleich eine Professionalisierung der Mitglieder einhergeht. Umgekehrt gilt diese Aussage ebenso (vgl. Rolff, 2007b, S. 14).

Bei dieser Betrachtung muss berücksichtigt werden, dass sehr wohl einzelne, isolierte Maßnahmen in den Bereichen der Personal- und Unterrichtsentwicklung initiiert werden können, welche keine Auswirkungen auf die jeweils anderen Komponenten nach sich ziehen. Diese Maßnahmen sind dann allerdings nicht auf die Komplexität der Organisation Schule bezogen und könnten daher dem Ziel einer umfassenden, nachhaltigen Schulentwicklung entgegenstehen (vgl. Horster & Rolff, 2006, S. 59).

Unerheblich davon, in welcher der drei Komponenten Maßnahmen forciert werden, vollzieht sich der Schulentwicklungsprozess in mehreren Phasen. Zunächst werden möglichst objektive Daten gesammelt bzw. erhoben, um den Ist-Zustand zu bewerten, Probleme zu erkennen und daraufhin Optimierungsbedarf zu ermitteln. Anschließend erfolgt ein Zielklärungsprozess, in welchem potentielle Lösungsansätze diskutiert und analysiert werden. Hierzu gehört auch die Überprüfung der Umsetzbarkeit der anvisierten Maßnahmen durch eine Beurteilung der zur Verfügung stehenden Ressourcen. Weiterführend werden die Entwicklungsvorhaben eingeführt und in der Praxis realisiert. Zuletzt schließt sich eine Evaluation an, anhand derer die Ergebnisse und Wirkungen beurteilt und der weitere Schulentwicklungsprozess geplant werden kann (vgl. Artelt, 2007, S. 132 f.; Rolff, 2007b, S. 150 f.).

Mithilfe der Schulentwicklung sollte die Einzelschule als eine lernende Institution begriffen werden, welche sich fortwährend um eine Qualitätssicherung und -verbesserung bemüht. Die Schule kann mehrere Lernebenen durchlaufen, wobei mit dem Erreichen einer höheren Lernebene die Kompetenz wächst, mit komplexen Situationen erfolgreich umgehen zu können (vgl. Grunder, 2004, S. 76). Die Mittel, welche diese Lernprozesse auslösen, stellen die Schulentwicklung dar. Daher ermöglicht es gerade die Schulentwicklung, dass sich die Ein-

zelschule im Rahmen des Programms der eigenverantwortlichen Schule profilieren und erfolgreich sowie nachhaltig arbeiten kann. Die Schulentwicklungsprozesse verlaufen hierbei nicht linear und können nicht von einzelnen wenigen Personen getragen werden kann. Sie bilden hingegen einen komplexen Systemzusammenhang. Schley (vgl. 2004, S. 22) führt als hauptsächliche hemmende Einflussfaktoren gelingender Schulentwicklung das Verharren im Entweder-oder-Denken an, welches kreative, synergetische Lösungen verhindere, sowie fehlende Teamstrukturen, unzureichende Ressourcennutzung und die Lokalisierung von Problemen außerhalb des eigenen Verantwortungsbereichs an.

5.3 Bedeutsamkeit der Evaluation für die Schulentwicklung

Ein zentrales Element zur Überprüfung der Wirksamkeit von Schulentwicklung bildet die Evaluation initiierter Maßnahmen. Klieme definiert Evaluation wie folgt:

> *Definition des Terminus „Evaluation" nach Klieme*
> „[Evaluation ist ein Prozess], innerhalb dessen eine zweckgerichtete Auswahl von Bewertungskriterien erfolgt, eine Institution oder Maßnahme auf Basis dieser Kriterien systematisch untersucht und bewertet wird und eine Kommunikation über die Bewertung stattfindet mit dem Ziel, Konsequenzen abzuleiten." (Klieme, 2005, S. 41).

Bezogen auf die Schulentwicklung umfasst Evaluation somit die systematische Überprüfung der Zweckmäßigkeit von Maßnahmen, welche der Qualitätsverbesserung dienen sollen (vgl. Posch & Altrichter, 1997, S. 29). Grundlage bilden möglichst objektive Datensammlungen, welche anschließend anhand festgelegter Kriterien ausgewertet werden. Auf Basis dieser Ergebnisse kann wiederum der Planungs- und Handlungsbedarf ermittelt werden, um den untersuchten Gegenstandsbereich zu verbessern oder zu optimieren. Folglich bilden die Resultate einer Evaluation neue Ausgangspunkte für die weitere Schulentwicklung (vgl. Holtappels, 2000, S. 43 ff.).
Wie bei der Schulentwicklung selbst findet mit der Evaluation gleichsam ein Lernprozess bei den betreffenden Personen statt. Aus diesem Grund fördert ein systematischer, bewusster Umgang mit den Evaluationsergebnissen die Professionalisierung innerhalb der Organisation Schule (vgl. Schley, 2004, S. 25) und liefert damit einen indirekten Beitrag zur Schulentwicklung. Evaluation dient zudem der Selbstvergewisserung und dem Anstoß neuer Projekte. Im Konzept der eigenverantwortlichen Schule kommt der Evaluation noch eine weitere

Funktion zu: der Rechenschaftslegung (vgl. Abschnitt 2.1). Die Schulen erhalten größere Freiheiten in Bezug auf die Gestaltung der Lernprozesse, müssen im Gegenzug jedoch die Qualität der stattfindenden Lernprozesse, meist gemessen an ihren Ergebnissen, nachweisen. Daher ist mit einer autonomer handelnden Schule zugleich die Forderung nach verstärkter Evaluation verbunden, deren Ergebnisse transparent zu kommunizieren sind.

Die verschiedenen Evaluationsformen können zum einen in Prozess- und Produktevaluation differenziert werden. Während bei der Prozessevaluation die Entwicklung selbst im Sinne einer formativen Messung (vgl. Abschnitt 3.2) im Zentrum der Untersuchung steht, wird bei letzterer lediglich das Ergebnis eines Lernprozesses in Form einer summativen Betrachtung beurteilt.

Zum anderen kann zwischen interner und externer Evaluation unterschieden werden. Die interne Evaluation ist ein schulinternes Verfahren unter Beteiligung der Schulmitglieder, in welchem vordergründig die Selbstreflexion und Selbstvergewisserung des eigenen pädagogischen Tuns steht, woraus sich wiederum Handlungsbedarf ableiten kann (vgl. Holtappels, 2000, S. 45). Bedeutend ist hierbei, dass eine interne Evaluation stets auch die Organisationsentwicklung anregt, da Kooperation und die Übernahme kollektiver Verantwortung als grundlegende Voraussetzungen für eine systematische interne Evaluation betrachtet werden (vgl. Peek, 2004a, S. 16). Die verschiedenen Modelle interner Evaluation lassen sich bei Holtappels (vgl. 2000, S. 45 ff.) nachlesen.

Wie bereits erläutert, ergibt sich die Notwendigkeit zur externen Evaluation bereits durch die Pflicht zur regelmäßigen Rechenschaftslegung. Eine Außensicht kann zudem helfen, den Ist-Zustand aus einem anderen Blickwinkel zu betrachten und auf diese Weise neue Perspektiven zu liefern. Kritische Reflexion, Bewahrung von Betriebsblindheit sowie die Feststellung von Stärken und Schwächen sind daher ebenfalls Funktionen der externen Evaluation (vgl. Artelt, 2007, S. 135). Daraus ergeben sich nach Artelt (vgl. 2007, S. 136) sechs Kriterien für eine gelingende externe Evaluation:

- Transparenz der Bewertungen und des angewandten Evaluationsverfahrens,
- verbindliche und geteilte Maßstäbe der Bewertung,
- Kontextsensitivität (Berücksichtigung der individuellen Rahmenbedingungen der Einzelschule),
- Ergebnisoffenheit der Evaluation,
- Klarheit und Verbindlichkeit des Evaluationsberichts sowie
- realistische Zielvereinbarungen als Folge der Evaluation.

Externe Evaluationen können in verschiedenen Formen erfolgen. Zum einen gibt es zentral administrierte standardisierte Tests, bei denen nicht während des Evaluationsverfahrens, sondern lediglich bei dessen Auswertung auf die spezifischen Bedingungen der Einzelschule Bezug genommen wird. Zum anderen sind systematische Expertenuntersuchungen wie die Schulinspektion in vielen Bundesländern mittlerweile fest etabliert. Es existieren des Weiteren Peer-Review-Verfahren, bei denen sogenannte „kritische Freunde", beispielsweise aus Nachbarschulen, die Evaluation durchführen, sowie Überprüfungen durch gemischte Teams aus internen wie externen Personen (vgl. Holtappels, 2000, S. 48 f.).

Interne und externe Evaluation stehen nicht komplementär zueinander, sondern sind gleichermaßen notwendige Elemente der Schulentwicklung. Laut Burkard bedürfe die interne Evaluation zur möglichst objektiven Bewertung der jeweiligen Situation, der erreichten Ergebnisse und zur Klärung der Beurteilungskriterien die Außensicht, so dass ein integriertes Modell von interner und externer Evaluation nötig sei (vgl. Burkard, 1995). Aus diesem Grund müsse nach Artelt (vgl. 2007, S. 137) eine Balance zwischen der Pflicht zur Rechenschaftslegung, intrinsischer Motivation zur Schulentwicklung und der Verantwortlichkeit gefunden werden. Maier (vgl. 2008a, S. 96 f.) führt demgegenüber an, dass in der Praxis das Verhältnis zwischen interner und externer Evaluation vielmals ungeklärt sei und eine Kopplung beider Instrumente oft nicht geschehe.

5.4 Einordnung der Verwendungsmöglichkeiten von Vergleichsarbeiten in die Schulentwicklungstheorie

Nachdem die Theorie der Schulentwicklung in Grundzügen dargestellt wurde, beschäftigen sich die Ausführungen in diesem Abschnitt damit, in welchen Bereichen der Schulentwicklung die Vergleichsarbeiten über die systematische Nutzung der Testrückmeldungen durch die Lehrkräfte und Schulleitungen einwirken können.

Zuvor sollte jedoch die Fragestellung problematisiert werden, inwiefern sich die Vergleichsarbeiten als eine interne bzw. als eine externe Evaluation typisieren lassen. Wie in Abschnitt 4.3.6.1 bereits erläutert wurde, enthalten die Tests mit ihren Rückmeldesystemen sowohl formative als auch summative Elemente. Hierbei muss jedoch zwischen den beiden Bereichen Anstoß von Entwicklungsprozessen sowie Evaluation klar differenziert werden. Die Ergebnisse der Vergleichsarbeiten können zwar für Schulentwicklungsabläufe genutzt werden, wie zum Beispiel für die Stärkung der diagnostischen Kompetenz der Lehrkraft. Dies entspricht dem formativen Charakter der Tests. Die Erhebung selbst ist jedoch von rein summativer Natur, da lediglich die Ergebnisse von Lernprozessen, gemessen an den bewer-

teten Schülerleistungen, ermittelt werden. Der Verlauf dieser Lernprozesse wird bei den Tests nicht berücksichtigt. Daher beinhalten die Vergleichsarbeiten als Evaluation eine ausschließlich summative Messung, die wiederum formative Entwicklungsprozesse anregen soll.

Bezüglich der Einordnung der Vergleichsarbeiten in die Kategorien interne oder externe Evaluation herrscht in der Fachliteratur bislang Uneinigkeit. Während Bohl, et al. (vgl. 2008, S. 462) und Leutner, et al. (vgl. 2007, S. 151) die Tests als ein verpflichtetes Instrument der Selbstevaluation auf Schul- und Klassenebene betrachten, typisiert Maag Merki (vgl. 2010, S. 146) die Vergleichsarbeiten als eine externe Evaluationsmethode. Letzteres erscheint zunächst logisch, da die Tests von einer äußeren Institution entwickelt und ausgewertet werden und sich daher dem Raster „zentrale administrierte standardisierte Tests“ zuordnen lassen. Jedoch müssten sie in diesem Fall auch den Kriterien von externer Evaluation genügen, die im Abschnitt 5.3 beschrieben wurden. Bei Betrachtung dieser Punkte ist zu erkennen, dass beispielsweise der Aspekt der Zielvereinbarungen auf die Vergleichsarbeiten nicht zutrifft. Die Rückmeldungen enthalten ausschließlich eine statistische Auswertung der Daten mit qualitativen Elementen der Diagnose. Es ist zwar durchaus möglich, dass Handlungsempfehlungen im Sinne einer Stärken-Schwächen-Analyse ausgesprochen werden; konkrete Zielvereinbarungen werden jedoch nicht getroffen. Die weitere Nutzung der Vergleichsarbeiten obliegt vielmehr ausschließlich der Verantwortung der Schulakteure.

Aus diesem Grund kann geschlussfolgert werden, dass mithilfe der Tests objektive Daten zu den Schülerleistungen erhoben und ausgewertet werden, die wiederum als Basismaterial für interne Evaluationen und die weitere Qualitätsentwicklung in der jeweiligen Schule genutzt werden können, aber nicht müssen. In dieser Betrachtung wären die Vergleichsarbeiten eher eine Komponente interner Evaluation. Inwiefern die betroffenen Lehrkräfte diese Sichtweise teilen, ist jedoch fraglich, da in der Mehrzahl der Bundesländer die Vergleichsarbeiten verpflichtend eingeführt wurden und sie somit der intrinsischen Motivation zur Selbstevaluation entgegenstehen können.

Das Hauptziel der Vergleichsarbeit liegt laut der EMSE im Anstoß von Schul- und Unterrichtsentwicklung (vgl. Netzwerk Empiriegestützte Schulentwicklung, 2006). Im Rahmen der Theorie von Schulentwicklung, wie sie in Grundzügen im Abschnitt 5.1 dargestellt wurde, ist Unterrichtsentwicklung jedoch eine Komponente von Schulentwicklung. Aus diesem Grund lässt sich die Aussage der EMSE bezüglich der Funktion von Vergleichsarbeiten auf den Anstoß von Schulentwicklung reduzieren. Organisations- und Personalentwicklung sind hierbei ebenfalls impliziert. Dies bedeutet nach Hofmann, et al. (vgl. 2005, S. 35), dass an den Schulen Verantwortlichkeiten, die Frage der Rechenschaftslegung und der Nutzung der Ergeb-

nisse in einem kommunikativen und kooperativen Prozess geklärt werden sollten. Für die Erläuterung der möglichen Nutzung der Vergleichsarbeiten für die Schulentwicklung ist eine Orientierung an den spezifischen Funktionen der Tests, wie sie in Abschnitt 4.2 ausgeführt wurden, hilfreich. Zum besseren Verständnis wird an dieser Stelle nochmals die grafische Darstellung eingefügt, welche die Vielzahl der mit den Vergleichsarbeiten verbunden Funktionen prägnant auflistet (vgl. Tabelle 3).

Funktionen von Vergleichsarbeiten			
Individual-diagnostik	**Schul-und Unterrichtsentwicklung**		**Bildungsmonitoring**
	Qualitätssicherung (summativ)	Qualitätsentwicklung (formativ)	
• Diagnostizieren und Fördern auf Schülerebene ?	• Schulevaluation • Bestandsaufnahme • Verortung • Rechenschaftslegung	• Innovationsanreiz für einen kompetenzorientierten Unterricht • Diagnostizieren und Fördern (auf Klassenebene) • Weiterentwicklung der Lehrerprofessionalität (Selbstreflexion, diagnostische Kompetenz, Orientierung an Bildungsstandards, Medienkompetenz) • Intensivierung der Fachgruppen- und Fachkonferenzarbeit	• Bestandsaufnahme • Feststellung von schulischem Unterstützungsbedarf • Unterstützung der Implementierung von Bildungsstandards • Transparenz der in den Bildungsstandards formulierten Anforderungen • Validierung der Kompetenzmodelle

Tabelle 3: Funktionen von Vergleichsarbeiten

Bei Betrachtung der Funktionen wird deutlich, dass die über die Rückmeldung gewonnenen summativen Informationen mittels einer Bestandsaufnahme, Verortung und Rechenschaftslegung dazu führen sollen, formative Prozesse der Schulentwicklung in Gang zu setzen. Ausgehend von den Funktionen der Vergleichsarbeiten wird in den folgenden Abschnitten der Versuch unternommen, aufzuzeigen, in welchen Bereichen der Schulentwicklung die Tests genutzt werden können bzw. inwiefern sie in Schulentwicklungsprozesse einwirken können. Hierbei wird im Sinne einer Strukturierung der Ausführungen eine Einteilung in die Komponenten Organisations-, Personal- und Unterrichtsentwicklung und somit eine Einordnung in die Schulentwicklungstheorie vorgenommen. Diese Zuordnung ist jedoch etwas einseitig, da Entwicklungsprozesse der einen Komponente automatisch Prozesse in den beiden anderen Dimensionen nach sich ziehen (vgl. Abschnitt 5.1). Aus diesem Grund besitzt die Einteilung der Nutzungsmöglichkeiten der Vergleichsarbeiten in die drei

Schulentwicklungskomponenten einen Modellcharakter, bei dessen Betrachtung der wechselseitige Systemzusammenhang stets berücksichtigt werden sollte.

5.4.1 Nutzung der Vergleichsarbeiten für die Organisationsentwicklung

5.4.1.1 Anstoß zur Strukturbildung und Kooperation

Bislang wurden Schulen vorrangig als Organisationen beschrieben, in denen die Lehrpersonen in ihrem Unterricht größtenteils als Einzelkämpfer eigenverantwortlich arbeiten. Diese pädagogische Freiheit verstärkte die Tendenz zur Individualisierung und Isolierung der Lehrerschaft (vgl. Münch, 2004, S. 28). Im Rahmen des Konzepts einer autonomer agierenden Schule, in der die Schulentwicklung eigenständig forciert und verwirklicht wird, werden jedoch andere Organisationsstrukturen benötigt, welche sich durch Kommunikation und Kooperation auszeichnen. Nach Hartung-Beck erfordern zudem die gegenwärtigen outputorientierten Reformen, wie die Bildungsstandards und Kerncurricula, „eine horizontale Integration aller Organisationsmitglieder […], d.h. es muss nicht mehr nur über Entscheidungen kommuniziert werden, sondern auch darüber, dass etwas entschieden werden muss“ (2009, S. 233). Grundvoraussetzung hierfür sei eine verstärkte Partizipation der Lehrerschaft in Entscheidungsprozesse mit kollektiv getragener Verantwortung. Auch Brinkmann-Hein & Reh (vgl. 2005, S. 30) betonen die zentrale Bedeutung von Kommunikationsnetzen für eine effektive Schulentwicklung. Es wird somit eine institutionell verankerte Infrastruktur benötigt, welche sowohl von der Schulleitung als auch vom Kollegium mitgetragen und von horizontalen und vertikalen Kommunikationskanälen betrieben wird (vgl. Hartung-Beck, 2009, S. 55; Rolff, 2007a, S. 41).

Die Kooperation kann sich auf verschiedene Bereiche erstrecken, indem sie zum einen auf eine Innovation bezogen und zum anderen personen- oder ergebnisorientiert sein kann. Gräsel, et al. (vgl. 2006, S. 209 ff.) unterscheiden zwischen drei verschiedenen Ausprägungsgraden der Kooperation:

1. *Austausch*

   Der Austausch umfasst eine wechselseitige Information, beispielsweise über Unterrichtsinhalte und -situationen. Auch die Weitergabe von Material an den Kollegen zählt hierzu. Da der Austausch Gelegenheitscharakter besitzt, arbeiten die Lehrpersonen weiterhin sehr autonom.

2. *Arbeitsteilige Kooperation*

   Bei dieser Form werden Aufgaben gezielt verteilt, um eine Arbeitserleichterung sowie eine Effizienzsteigerung zu bewirken. Dennoch findet kein gemeinsames Arbeiten im wörtlichen Sinne statt, da die Ausführung der Aufgaben weiterhin individuell verläuft und nur die Zielsetzung und das Ergebnis abgestimmt werden.

3. *Kokonstruktion*

   Bei der Kokonstruktion finden ein gemeinsames Lernen und die Entwicklung kollektiv getragener Problemlösungen statt. Der Arbeitsprozess erfolgt größtenteils in Zusammenarbeit, so dass die Autonomie der Einzellehrkraft deutlich reduziert wird. Verbunden ist dies mit dem Effekt der individuellen Professionsentwicklung, indem ein Lernen voneinander stattfindet. Die Kokonstruktion ist somit die stärkste Form von Kooperation.

Die vorgestellten Kooperationsausprägungen lassen sich ähnlich wie ein Kompetenzniveaumodell beschreiben, indem die Kooperation in Form einer kontinuierlichen Kompetenzentwicklung stetig gesteigert und intensiviert werden kann (vgl. Schweizer & Klieme, 2005). Die Form der Kokonstruktion wird in der Fachliteratur auch als „professionelle Lerngemeinschaft“ thematisiert (vgl. Reh, 2008, S. 167; Rolff, 2007b, S. 114 ff.). Grundbedingungen für jegliche Art von Kooperation sind gemeinsame Zielsetzungen, Vertrauen zu den Handlungen anderer, welche sich der eigenen Kontrolle entziehen, sowie die Bewahrung einer gewissen Handlungs- und Entscheidungsfreiheit (vgl. Gräsel, Fußangel, & Pröbstel, 2006, S. 207 f.).

Die Vergleichsarbeiten können als Anstoß dienen, kommunikative und kooperative Strukturen institutionell in der Schulorganisation zu initiieren bzw. bereits bestehende zu festigen. Gemeint ist hiermit der Aufbau einer Infrastruktur, in welcher gegenseitiges Feedback und die Rückmeldeergebnisse aus den Tests als zentrale Anhaltspunkte für die weitere Steuerung verwendet werden. Als primäre Adressaten werden neben den Lehrkräften die Fachgruppen und die Fachkonferenzen genannt, so dass die Auswertung der Ergebnisse nach Aussage von Peek, et al. (vgl. 2006, S. 231) in einem gemeinsamen Analyseprozess erfolgen sollte, aus dem wiederum Konsequenzen für die Schulentwicklung gezogen werden. Hierfür ist ein strukturiertes, systematisches Vorgehen erforderlich, indem beispielsweise feste Arbeitszeiträume für größere Verbindlichkeit sorgen und die jeweiligen Konferenzen thematisch vorbereitet werden. Ebenso wird die Ernennung von Koordinatoren für die Vergleichsarbeiten empfohlen (vgl. Schneewind, 2007b, S. 85), welche sowohl den organisatorischen Ablauf der Tests regeln als auch als Ansprechpartner innerhalb der Schule dienen.

Im Sinne einer Verankerung von Teamstrukturen ist die Bildung von temporären Arbeitsgemeinschaften der an den Tests teilnehmenden Lehrkräfte denkbar, deren Organisator oder Vorsitzender der Koordinator wäre. In einer solchen Zusammenarbeit könnten die Analyse- und Auswertungsprozesse gemeinsam erfolgen. Die Ergebnisse dieser Kooperation könnten wiederum für das übrige Kollegium aufbereitet und zum Beispiel in Gesamtkonferenzen vorgestellt werden, so dass die Arbeitsgruppe als Multiplikator der Erkenntnisse aus den Vergleichsarbeiten dienen würde und den Handlungsbedarf in den jeweiligen Bereichen zumindest benennen bzw. Maßnahmen hierzu anstoßen könnte. Die Funktionen dieser auf die Vergleichsarbeiten bezogenen Arbeitsgemeinschaft wären daher ähnlich denen einer Steuergruppe. Der Vorteil einer solchen Struktur liegt darin, dass die Mitglieder dieser Gruppe bei jedem Testdurchlauf wechseln und ein Großteil des Kollegiums auf diese Weise langfristig involviert sein würde. Bei fachspezifischen Besprechungen wäre es sinnvoll, die Teamstruktur fächerweise zu splitten und die Partizipation des jeweiligen Fachsprechers oder Aufgabenfeldleiters zu befördern. Die Zusammenarbeit würde im Sinne dieser Arbeitsgemeinschaft wenigstens auf der zweiten Stufe, der arbeitsteiligen Kooperation, verlaufen und vermehrt auch Anreize für eine Kokonstruktion bieten.

Die Inhalte einer solchen kooperativen Nutzung der Vergleichsarbeiten können verschiedener Natur sein. Da in den Rückmeldungen die Leistungen einzelner Klassen wiedergespiegelt werden, betreffen die Ergebnisse aus den Teamsitzungen vermutlich primär die Unterrichtsentwicklung. Positive Erfahrungen aus dieser Zusammenarbeit können allerdings Effekte auf die generelle Kooperationsbereitschaft bewirken, so dass in dessen Folge auch in anderen thematischen Feldern Teamstrukturen intensiviert werden würden (vgl. ebd., S. 86).

Die Vergleichsarbeiten können des Weiteren auch als Anstoß für die Öffnung der Schule nach außen fungieren, indem externe Netzwerke stabilisiert werden. In diesem Zusammenhang ist exemplarisch betrachtet ein regionaler Qualitätszirkel denkbar, in welchem Erfahrungen und Anregungen bezüglich der Testnutzung ausgetauscht werden können. Dies würde jedoch eine professionelle Moderation und administrative Unterstützung erfordern (vgl. Haenisch & Müller, 2005, S. 308).

Die Effekte von Kooperation sind generell nicht zu unterschätzen. Sie wirken in die Personalentwicklung hinein, indem sie nach Reh (vgl. 2008, S. 163) Anreize für die Weiterentwicklung der Selbstreflexionsfähigkeit der Lehrkräfte setzen und die Kooperation als Arbeitserleichterung empfunden werden kann. Auch würde durch die kollektiv ausgeübte Verantwortung der Druck hierarchischer Strukturen herabgesetzt werden und dennoch die Eigenverantwortlichkeit für die Lernprozesse im Unterricht bestehen bleiben (vgl. ebd., S.

166). Zugleich kann Kooperation als ein Instrument betrachtet werden, um Demokratisierung innerhalb der schulischen Organisationsstruktur zu befördern (vgl. Brinkmann-Hein & Reh, 2005, S. 30). Gräsel, et al. (vgl. 2006, S. 205) betrachten Kooperation zudem als eine fördernde Bedingung, um Innovationen einzuführen und umzusetzen. Daher sei anzunehmen, dass besonders erfolgreiche Schulen von einem hohen Maß an Kooperation im Kollegium charakterisiert seien.

Obwohl die Zusammenarbeit von den Schulakteuren gewünscht wird, ist deren Ausmaß im Schulalltag jedoch relativ gering (vgl. Gehrmann, 2003, S. 274, 287; Ulich, 1996, S. 150 ff.). Während nach Aussage von Steinert, et al. (vgl. 2006, S. 200) die Schulform kein entscheidender Faktor für den Kooperationsgrad sei, erläutern Gräsel, et al. (vgl. 2006, S. 207), dass aufgrund unterschiedlicher Rahmenbedingungen die Zusammenarbeit an Gymnasien am geringsten ausgeprägt sei. Hinzu kommt, dass die Kooperation umso problematischer wird, je stärker der eigene Unterricht Gegenstand der Zusammenarbeit ist (vgl. Reh, 2008, S. 165). Dies sollte auch in Bezug auf die Vergleichsarbeiten nicht unberücksichtigt bleiben, weil hier die Ergebnisse der eigenen Schüler sowie die weitere unterrichtliche Tätigkeit im Mittelpunkt stehen.

Verschärft wird dieser Aspekt durch den sozialen Vergleich, bei dem die Leistungen der eigenen Klasse an den Ergebnissen der Parallelklassen oder dem korrigierten Landesmittelwerts gemessen werden. Aus diesem Grund ist es durchaus möglich, dass sich eine Teamstruktur zur Thematik Vergleichsarbeiten nicht von selbst entwickelt, sondern zuerst von einer Person oder Personengruppe, wie der erweiterten Schulleitung oder dem Koordinator, angestoßen und forciert werden muss. Zugleich ist es wichtig, dass die Kooperation nicht ohne Einwilligung der betreffenden Personen verordnet wird, sondern sich intrinsisch motiviert entwickelt (vgl. Grunder, 2004, S. 85). Erzwungene Teamarbeit wird hingegen als eine Einschränkung des eigenen Handlungsspielraumes und der Autonomie wahrgenommen werden. Dies stellt zugleich eine generelle Problematik in der Nutzung der Vergleichsarbeiten dar: Die Durchführung der Tests ist (teilweise) verpflichtend, die Nutzung obliegt jedoch der Eigeninitiative der jeweiligen Lehrperson. Um die Ergebnisse nicht nur für sich selbst zu analysieren und zu verwenden, sondern in einer Kooperation mit den anderen Lehrpersonen, bedarf es nochmals einer zusätzlichen Motivation.

5.4.1.2 Funktionen der Schulleitung

Die Schulleitung nimmt innerhalb der Organisationsstruktur der Schule eine besondere Stellung ein, da sie als hierarchisches Element über Entscheidungsbefugnisse in den Bereichen der Verwaltung und der Gestaltung der Organisation verfügt (vgl. Buchen, 2009, S. 38; Kuper, 2008, S. 159). Sie trägt die Gesamtverantwortung für das Funktionieren der Arbeitsprozesse und für die Erfüllung des Bildungs- und Erziehungsauftrages (vgl. Orth, 2001, S. 220). Das Aufgabenprofil der Schulleitung beeinflusst insbesondere die Organisationsentwicklung. Hierzu zählen beispielsweise die sachgerechte Umsetzung von administrativen Vorschriften und Konferenzbeschlüssen, die Organisation des Betriebsablaufs sowie der Aufbau einer innerschulischen Infrastruktur, in welcher kollektive Entscheidungen getroffen werden können (vgl. Bonsen, 2010, S. 277; Kuper, 2008, S. 160). Im Kontext einer zunehmend eigenverantwortlich agierenden Schule erweitern sich die Funktionen der Schulleitung unter anderem um die Entscheidungskompetenz für den systematisch strukturierten Ressourceneinsatz, sowohl in finanzieller als auch in personeller Hinsicht. Dies greift insbesondere in die Personalentwicklung ein, indem die Schulleitung mehr Mitbestimmungsrechte im Bereich der Personalauswahl erhält und die Beratungs- und Fortbildungsmaßnahmen gezielt steuern kann (vgl. Bonsen, 2010, S. 280 f.). Als Schlüsselaufgaben einer Schulleitung nennt Malik (vgl. 2001, S. 171 ff.) daher folgende Aspekte: Für Ziele sorgen, organisieren, entscheiden, kontrollieren, messen, beurteilen sowie die Selbstentwicklung von Menschen fördern und unterstützen.

Obwohl die Schulleitung in das Kerngeschäft der Schule, dem Unterricht, nur indirekt einwirken kann (vgl. Hallinger & Heck, 1995, S. 732), ist sie dennoch für die Sicherung und Entwicklung der Unterrichtsqualität verantwortlich. Daher sollte die Leitung im Sinne einer nachhaltigen Schulentwicklung darin bestrebt sein, Qualitäts- und Evaluationsprogramme zu verankern (vgl. Bonsen, 2010, S. 280 f.) und intensiv die Organisations-, Personal- und Unterrichtsentwicklung zu befördern. Umgekehrt gilt dies ebenso: „Umfassende Schulentwicklungsprozesse ohne aktive Unterstützung der Schulleitung sind undenkbar“ (Rolff, 2007a, S. 51). Dies impliziert, dass vom Kollegium initiierte Maßnahmen und Innovationen nicht gegen den Willen der Schulleitung umgesetzt werden können. Der Schulleitung obliegt zudem die Funktion der Förderung von Veränderungen (vgl. Münch, 2004, S. 26), zu denen auch äußere Reformkonzepte wie die Bildungsstandards oder die Vergleichsarbeiten zählen. Hierfür sind das professionelle Selbstverständnis der Schulleitung und deren persönliche Einstellung Innovationen gegenüber entscheidend. Erkenntnisse der Untersuchung von Bonsen, et al. (vgl. 2002, S. 115 ff.) ergaben, dass das Kollegium umso innovationsfreudiger ist, je zielbezogener die Schulleitung arbeitet. Folglich kann sich die Bereitschaft der

Schulleitung zur Teilnahme an den Vergleichsarbeiten sowie das damit verbundene Engagement positiv auf die Akzeptanz der Tests im Kollegium auswirken. Negative Effekte sind jedoch ebenfalls möglich: Wenn die Schulleitung das Konzept der Vergleichsarbeiten für die Schulentwicklung nicht als förderlich einschätzt, kann dies zu einer zurückhaltenden Einstellung bei den betreffenden Lehrkräften führen, wodurch im Rückschluss wiederum eine intensive Nutzung der Rückmeldungen erheblich erschwert wird. Demnach nimmt die Schulleitung bezüglich der Akzeptanz und Nutzung der Vergleichsarbeiten bei den Lehrpersonen eine nicht zu unterschätzende Schlüsselposition ein!

Die Schulleitung kann in diesem Zusammenhang mehrere Maßnahmen ergreifen, welche die Bereitschaft der Lehrer mit den Tests zu arbeiten erhöhen können. Zum einen sollten Informationen über das Testkonzept bereitgestellt und der Zusammenhang zwischen den Vergleichsarbeiten und den Bildungsstandards hinreichend erläutert werden. Dies sorgt für Transparenz und Aufklärung. Des Weiteren kann die Schulleitung eine Art Vorbildfunktion einnehmen, indem deren Mitglieder ebenfalls an den Vergleichsarbeiten in ihren Klassen teilnehmen und dem Kollegium ihre Erfahrungen kommunizieren und auf diese Weise motivierend wirken können (vgl. Haenisch & Müller, 2005, S. 310). Zudem sollte die Schulleitung die Schaffung von Arbeitsstrukturen befördern, welche eine kooperative Auseinandersetzung und Interpretation der Testergebnisse ermöglichen (vgl. Abschnitt 5.4.1.1). Da die Durchführung, Korrektur und Dateneingabe einen erheblichen Arbeitsaufwand für die Lehrpersonen bedeuten, obliegt es letztlich der Schulleitung, nach Lösungsansätzen zu suchen, um Zeitressourcen oder Arbeitserleichterung zu schaffen, so dass das Engagement der Lehrkräfte dauerhaft bestehen bleibt (vgl. Rolff, 2007a, S. 51).

Die Schulleitung kann die Ergebnisse der an den Tests teilnehmenden Klassen auch für sich effektiv nutzen, da mithilfe des sozialen Vergleichs mit dem korrigierten Landesmittelwert eine Verortung des Outputs an ihrer Schule erreicht wird. Hieraus können wiederum strategische Zielsetzungen und Schulentwicklungsmaßnahmen abgeleitet werden, um die Qualität langfristig zu steigern (vgl. Bonsen, 2010, S. 288). Beispielsweise ist es möglich, dass die Erkenntnisse aus den Vergleichsarbeiten zu weiteren schulinternen Evaluationen anregen und neue Entwicklungsschwerpunkte setzen, welche im Schulprogramm dokumentiert werden können. Denkbar wäre auch, die Ergebnisse als Rechenschaftslegung nach außen zu benutzen, so dass die nachgewiesenen positiv bewerteten Schülerleistungen zur Profilierung dienen. Dies ist allerdings nicht im Sinne der Zielsetzung der Vergleichsarbeiten und kann zu einem inoffiziellen Ranking der Schulen einer Region führen.

5.4.2 Nutzung der Vergleichsarbeiten für die Personalentwicklung

5.4.2.1 Stärkung des professionellen Selbst

Die Vergleichsarbeiten können des Weiteren im Bereich der Personalentwicklung genutzt werden, indem mittels objektiver Daten über die Leistungen der Schüler das eigene unterrichtliche Handeln überprüft wird und auf diese Weise die Professionalisierung der Lehrkräfte gestärkt werden kann (vgl. Hartung-Beck, 2009, S. 27). Dazu muss zunächst betrachtet werden, inwiefern die Tests als ein Diagnose- und Evaluationsinstrument den Professionalisierungsgrad beeinflussen können. Bauer (vgl. 2009a, S. 75 ff.) führt in diesem Zusammenhang begründend an, dass über Evaluationen – und somit auch über die Vergleichsarbeiten – die Lehrkräfte mit der Wirksamkeit des eigenen Handelns konfrontiert würden, was wiederum Prozesse der Selbstentwicklung und Handlungsoptimierung anregen könne. Dies stellt oftmals eine neue Erfahrung für die Lehrerschaft dar, welche mit dem Umgang mit Feedback nicht vertraut ist. Die Lehrpersonen erhielten bislang die zentralen Informationen über die Qualität ihres Handelns über den Unterricht selbst, zum Beispiel in Form von Klassenarbeiten (vgl. Bauer, 2009b, S. 227). Diese Art der Selbstreflexion bzw. Selbstevaluation, welche in unterschiedlichem Ausmaß professionell gesteuert sein kann, ist ein bedeutsames Element der eigenen Kompetenzentwicklung und wird nun durch die externe Komponente der Vergleichsarbeiten ergänzt. „Nur wer fortlaufend überprüft, wo er steht, was er oder sie erreicht hat und was nicht, kann sein Lernen selber steuern, bleibt auf Dauer überhaupt lernfähig“ (Rolff, 2007b, S. 145).

Hierzu erläutert Bauer (vgl. 2009a, S. 75), dass Evaluationen nur dann zu einer weiteren Professionalisierung und einer qualitativen Handlungsverbesserung führen würden, wenn die Evaluation von der jeweiligen Person als bedeutsam empfunden und systematisch mit dem Ziel der Veränderung verwendet werde. Insbesondere bei externer Rückmeldung wie bei den Vergleichsarbeiten ist die Akzeptanz stark davon abhängig, ob die Kriterien zur Bewertung von dem Adressaten als adäquat betrachtet werden (vgl. Schley, 2004, S. 21). Bisherige Untersuchungen zum Umgang mit Evaluationen ergaben, dass Lehrkräfte mit zuvor gesammelten Evaluationserfahrungen mit einer höheren Wahrscheinlichkeit dazu bereit sind, die Ergebnisse zu nutzen (vgl. Reh, 2008, S. 163). Dies ist insbesondere für die Vergleichsarbeiten ein bedeutsamer Fakt, da die Tests jährlich durchgeführt werden und die Fachschaften sich somit regelmäßig mit dem Instrument konfrontiert sehen. Die Kommunikation bisheriger Erfahrungen kann daraus schlussfolgernd für eine Nutzung der Rückmeldungen förderlich sein, insofern diese Erfahrungen selbst als positiv beurteilt wurden.

Wie die Lehrpersonen die Tests letztlich nutzen, hängt zu einem hohen Ausmaß von ihrem professionellen Selbst ab. Es entwickelt sich prozessartig durch die Ausbildung, Fortbildung, Trainings und praktischen Erfahrungen. Eine zentrale Funktion nehmen die systematische Reflexion und Evaluation als Voraussetzungen für fortwährendes Lernen und optimiertes Handeln ein (vgl. Bauer, 2009a, S. 83; Rolff, 2007b, S. 148). Das theoretische Konstrukt des professionellen Selbst besteht einerseits aus pädagogischen Basiskompetenzen und Fachkompetenzen, welche die Grundbedingungen für wirksames Unterrichtshandeln darstellen. Andererseits zählt hierzu ein internes Steuerungsinstrument, das individuelle Werte, Ziele und Annahmen über das eigene pädagogische Kompetenzprofil und die Selbstwirksamkeit umfasst (vgl. Bauer, 2009a, S. 84 f.). Professionalisierung verlangt demnach nach einer selbstreflexiven Haltung. Dies setzt jedoch auch eine gewisse Offenheit und Ehrlichkeit gegenüber sich selbst voraus (vgl. Rolff, 2007b, S. 146).

Die Vergleichsarbeiten können die Professionalisierung weiterentwickeln, indem eine Interpretation der Ergebnisse zugleich eine Selbstreflexion über das eigene Agieren bedingt. Zum einen müssen die Lehrpersonen somit fähig sein, aufgrund der Klassendaten aus den Tests fundierte Rückschlüsse über die Qualität des absolvierten Unterrichts rückblickend zu ziehen (vgl. Bauer, 2009b, S. 225). Hierfür benötigen die Lehrenden sowohl theoretisches Wissen über die Bedingungen von Unterrichtserfolg als auch die Kompetenz, dieses Wissen praktisch anzuwenden und Problemlösungen zu finden. Zum anderen können die Vergleichsarbeiten nur dann nachhaltige Effekte auslösen, wenn diese Erkenntnisse auch auf zukünftige Handlungen angewandt werden und somit das unterrichtliche Tun qualitativ verbessert wird. Diese Form der Professionalisierung ist ebenfalls als eine Kompetenzentwicklung zu verstehen, indem Prozesse und Ergebnisse miteinander in Beziehung gesetzt werden und anschließend eine Handlungskompetenz entsteht (vgl. Bauer, 2009a, S. 91; Kuper & Hartung, 2007, S. 217; Oelkers, 2009, S. 3).

Aber auch weitere Kompetenzen der Lehrpersonen können mithilfe der Vergleichsarbeiten gefördert werden. Die internetbasierte Dateneingabe und die Bereitstellung der Rückmeldung erfordern ein gewisses Maß an Medienkompetenz. Die intensive Auseinandersetzung mit der Rückmeldung schult zudem statistische Kenntnisse und das Evaluationswissen (vgl. Kiper, 2009, S. 20 f.), da beispielswiese Informationen unterschiedlicher Aggregationsniveaus bewertet werden müssen (vgl. Kuper, 2008, S. 161). Dies kann wiederum gewinnbringend in Formen der Selbstevaluation münden, bei denen Daten ausgewertet und interpretiert sowie Maßnahmen abgeleitet werden müssen (vgl. Bauer, 2009a, S. 87).

5.4.2.2 Entwicklung von diagnostischen Kompetenzen

Neben dem professionellen Selbst und der Evaluationskompetenz kann zudem die diagnostische Kompetenz der Lehrkräfte mithilfe der Vergleichsarbeiten ausgebaut und gefestigt werden. Nach Weinert, et al. (1990) zählt die diagnostische Kompetenz zu den vier Schlüsselkompetenzen für eine professionelle Berufsausübung als Lehrperson. Demnach ist die diagnostische Kompetenz eine zentrale Bedingung für erfolgreichen Unterricht und beeinflusst in hohem Maße die Unterrichtsgestaltung sowie den Unterrichtserfolg (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 128).

Sie besteht dabei aus einem Bündel an Fähigkeiten, um „den Kenntnisstand, die Lernfortschritte und die Leistungsprobleme der einzelnen Schüler im Unterricht fortlaufend beurteilen zu können, so dass das pädagogische Handeln auf diagnostischen Einsichten aufgebaut werden kann“ (Lorenz, 2005, S. 318). Grundlage für die Diagnosefähigkeit sind die von der Lehrperson getroffenen Einschätzungen über eine Merkmalsausprägung. Die Diagnosegenauigkeit wiederum umfasst das Maß der Übereinstimmung dieser Annahmen mit der tatsächlichen Merkmalsausprägung (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 120).

Die diagnostische Kompetenz entwickelt sich nach Helmke, et al. (vgl. 2004, S. 121) über schwer beeinflussbare Merkmale, wie Intelligenz, kognitive Fähigkeiten, erfahrungsbedingte Wissensstrukturen und Fertigkeiten. Hierzu zählen sowohl methodisches Wissen über Beurteilungsformen, gegenstandsspezifisches Wissen über die Anforderungen von Lerninhalten, Merkmale von Aufgabenschwierigkeiten und Lösungsstrategien zur Bewältigung von Aufgaben, als auch spezifisches schülerbezogenes Wissen über die Lerngruppe, wie deren Motivation (vgl. ebd., S. 121). Eine ausgeprägte Diagnosefähigkeit zeichnet sich durch das Bemühen um möglichst objektive Einschätzungen, gestützt auf spezifischen Beobachtungen und das Wissen über Bewertungsmethoden und -fehler, aus. Hierzu zählt ebenso die selbstreflexive Bereitschaft, die eigenen Urteile kontinuierlich zu hinterfragen und zu bewerten (vgl. Tresch, 2007, S. 122 f.).

Annahmen über die Leistungsfähigkeit einzelner Schüler können anhand von drei Komponenten vorgenommen werden: Die erste Komponente ist die Niveaukomponente, mit welcher ermittelt werden kann, ob die Lehrperson die Leistungsfähigkeit im Mittel zu hoch, gerade richtig oder zu niedrig einschätzt. Eventuelle systematische Fehleinschätzungen können hierbei ebenfalls tendenziell sichtbar werden. Zudem ist die Streuungskomponente zu nennen, bei welcher die Diagnosegenauigkeit erfasst wird. Zuletzt gibt es die Rangordnungskomponente, welche die Fähigkeit ausdrückt, Kompetenzabstufungen zwischen Schülern einzuschätzen oder Aufgabenschwierigkeiten in einer Rangfolge zu modellieren (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 123 f.).

Die Lehrpersonen können sich bei der Ausübung ihrer diagnostischen Fähigkeiten auf die bereits erläuterten Bezugsnormen stützen (vgl. Abschnitt 4.3.6.2). Zum einen können sie in Form eines sozialen Vergleichs die Leistungen der Schüler einer Lerngruppe untereinander einschätzen. Zum anderen kann die kriteriale Bezugsnorm verwendet werden, indem die Schülerleistung anhand eines sachlichen Bezugs, wie zu den Kompetenzstufen oder den Anforderungen der Lerninhalte, beurteilt wird. Zuletzt können ipsative Vergleiche erstellt werden, bei welchen die individuelle Lernentwicklung innerhalb einer bestimmten Zeitspanne betrachtet wird (vgl. ebd., S. 124 f.).

Die Vergleichsarbeiten stellen nun eine Gelegenheit dar, die eigene diagnostische Kompetenz zu überprüfen und weiterzuentwickeln, da die Lehrkräfte über die Rückmeldung objektive Daten in Form einer Stärken-Schwächen-Analyse über die Leistungsfähigkeit ihrer Lerngruppe in einigen fachlichen Teilbereichen erhalten. Beispielsweise können sie die eigenen vorherigen Einschätzungen zu den Aufgabenschwierigkeiten mit den späteren Ergebnissen abgleichen. In diesem Bereich kann analysiert werden, welche Items zu schwer oder zu leicht für die Schüler waren und worin die Gründe hierfür zu sehen sind. Es ist des Weiteren notwendig zu wissen, welche didaktischen Modelle den Aufgaben zugrunde liegen, welche Lösungsschritte notwendig sind und welche Fehlerquellen auftreten können. Auf diese Weise werden didaktische Überlegungen darüber angeregt, mit welchen Schwierigkeitsgraden die Lernenden in künftigen Unterrichtssequenzen konfrontiert werden können, so dass eine Anpassung der Ergebnisse mit den bisher verwendeten Aufgaben erfolgen kann (vgl. Helmke A. , 2007, S. 225; Helmke, Hosenfeld, & Schrader, 2004, S. 121).

Weiterhin ist es möglich, die für den Test prognostizierten Ergebnisse mit den tatsächlichen Daten im Nachhinein zu vergleichen. Über- und Unterschätzungen können dabei auffällig werden. In diesem Kontext wird die Selbstreflexion über die Hintergründe des eigenen Urteils angeregt. Hierzu zählt eine Beleuchtung der möglichen Einflussfaktoren auf die Ergebnisse, wie zum Beispiel die Motivation der Schüler oder inwiefern die getesteten Kompetenzen tatsächlich im Unterricht gefördert wurden. Bedeutsam ist ebenfalls, ob die Schüler an die Itemformate gewöhnt sind (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 138 f.). Zuletzt sollte die betreffende Lehrkraft versuchen, die Ursachen für eventuelle eigene Urteilsfehler zu erfragen. Eine typische Fehlerquelle, die zu Bewertungsverzerrungen führt, ist der Erwartungseffekt, auch Pygmalion-Effekt genannt, bei welchem die eigene Erwartungshaltung an die Leistung der Schüler stärker im Vordergrund steht als die tatsächliche Leistungsfähigkeit. Des Weiteren ist die Tendenz zur Mitte zu erwähnen, bei der sehr gute oder sehr schlechte Einschätzungen vermieden werden, so dass die gesamte Bewertungsskala nicht ausgenutzt wird. Stattdessen konzentrieren sich die Urteile im mittleren Leistungs-

spektrum. Zuletzt ist auch der Ausstrahlungseffekt (Halo-Effekt) bedeutsam, bei dem zu Pauschalurteilen tendiert wird beziehungsweise Vorurteile die Bewertung beeinflussen (vgl. Tresch, 2007, S. 121).

In einigen Bundesländern wird die Überprüfung der eigenen Diagnosekompetenz von den Landesinstituten verstärkt forciert, indem die Lehrkräfte vor der Durchführung der Vergleichsarbeiten dazu aufgefordert werden, anhand einer Aufgabenanalyse eine Prognose über das Abschneiden ihrer Lerngruppe bei den jeweiligen Items abzugeben (vgl. Hosenfeld, 2005, S. 113). Dieses Vorgehen hat den bedeutsamen Vorteil, dass die Lehrpersonen nicht nur zu einer intensiven Auseinandersetzung mit dem Testgegenstand animiert werden, sondern ihre eigene Einschätzung schriftlich erhalten und der Vergleich mit den tatsächlichen Leistungsergebnissen somit vereinfacht wird. Die Überprüfung der eigenen Urteile und die intensive Auseinandersetzung mit den Rückmeldeinformationen können zu einer Stärkung und Weiterentwicklung der diagnostischen Kompetenz beitragen (vgl. Tresch, 2007, S. 120). Allerdings müssen die daraus gewonnenen Erkenntnisse auf das unterrichtliche Agieren übertragen werden in Form einer gezielten Anpassung der Lernsituationen an den Leistungsstand einzelner Schülergruppen (vgl. Helmke, Hosenfeld, & Schrader, 2004, S. 128 f.). Auf diese Weise kann die Unterrichtsqualität verbessert werden, was indirekt den Erfolg der Lernprozesse beeinflusst.

5.4.3 Nutzung der Vergleichsarbeiten für die Unterrichtsentwicklung

Die Implementierung der Vergleichsarbeiten erfolgte in Hinblick auf Standardisierung und Vergleichbarkeit. Letztlich ist es das Ziel der aktuellen Reformmaßnahmen, die derzeitige Bildungsqualität zu steigern. Da die individuellen Lernprozesse im Unterricht stattfinden, wird insbesondere auf dieser Ebene eine Qualitätsentwicklung angestrebt. Mittels der Vergleichsarbeiten kann der Erfolg und die nachhaltige Wirkung der Lernsituationen überprüft werden. Daher erscheint es als eine Selbstverständlichkeit, dass sich die Tests mit den jeweiligen Rückmeldungen auch für die Unterrichtsentwicklung nutzen lassen sollten. Voraussetzung hierfür ist die tiefgründige Analyse der Schülerergebnisse, aus welcher die Konsequenzen und der Handlungsbedarf abgeleitet werden müssen. Diese Auseinandersetzung sollte, wie bereits beschrieben, möglichst in Kooperation mit den anderen betreffenden Lehrpersonen erfolgen (vgl. Abschnitt 5.4.1.1). Die Umsetzung und inhaltliche Ausgestaltung der vereinbarten Maßnahmen obliegt jedoch der Professionalität der individuellen Lehrkraft und verläuft im Rahmen der ihr zur Verfügung stehenden pädagogischen Freiheit. Die hierbei angestrebten Ziele sollten entsprechend des SMART-Prinzips spezifisch, mess-

bar, attraktiv sowohl für den Lehrenden als auch die Lernenden, realistisch und terminiert sein (vgl. Tresch, 2007, S. 134 f.).

Die aus den Ergebnissen abgeleiteten Maßnahmen können kurz-, mittel- oder langfristig angelegt sein. Generell können die Vergleichsarbeiten in die verschiedensten Bereiche hineinwirken, so dass an dieser Stelle keine vollständige Auflistung möglich ist. Es ist vielmehr bedeutsam, dass die Lehrkräfte bewusst und gezielt Konsequenzen einleiten und sich nachdrücklich an den ausgewiesenen Schwachstellen ihrer Lerngruppe orientieren, um somit eine Leistungsverbesserung zu erzielen. Dennoch sollen im Folgenden die einzelnen Bereiche der Unterrichtsentwicklung erwähnt werden, die durch die Vergleichsarbeiten beeinflusst werden können (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 97 f.).

Zum einen können die *Unterrichtsinhalte und Kompetenzbereiche* Gegenstand der Überlegungen werden. Falls aus der Rückmeldung ersichtlich wird, dass einzelne Schülergruppen Defizite in bestimmten inhaltlichen Lernfeldern aufweisen, können zum Beispiel die betreffenden Thematiken wiederholt und vertieft bzw. einzelne Kompetenzbereiche nochmals gezielt gefördert werden. Dies kann sich auch langfristig auf die Unterrichtsplanung in den Folgejahren auswirken, indem eine veränderte Schwerpunktsetzung erfolgt.

Des Weiteren können die *Materialien* angepasst und neue Aufgabenformate oder Aufgabenstellungen entwickelt werden. Hierbei ist eine Orientierung an den Items der Vergleichsarbeiten möglich. Die reine Übernahme von Multiple-Choice-Formaten wird allerdings nicht die Unterrichtsqualität erhöhen. Vielmehr sollten die Materialien aus den Tests durch Abwandlung in Lernaufgaben dazu anregen, ebenfalls kompetenzorientierte Übungen verstärkt in den Unterricht zu integrieren.

Drittens kann die *Unterrichtsplanung und -gestaltung* mit dem zugrunde liegenden didaktischen Verständnis weiterentwickelt und optimiert werden. Dabei kommt den Vergleichsarbeiten die zentrale Funktion zu, die Lehrkräfte stärker an das Konzept der Bildungsstandards heranzuführen, so dass sich der Unterricht zunehmend an dem kompetenzorientierten Gestalten von Lernsituationen orientiert beziehungsweise die Lehrenden in ihrer bisherigen Vorgehensweise dahingehend bestärkt werden. Durch die Tests sollten sie einen Einblick erhalten, wie die Kompetenzbereiche und -dimensionen aufgebaut sind und auf welche Weise erworbene Kompetenzen überprüft werden können. Die Schwierigkeit liegt jedoch in der Anwendung dieser Erkenntnisse in Hinblick auf die Frage, wie sie kompetenzorientierte Lernvorgänge selbst initiieren und gestalten können. Zu diesem Aspekt liefern die Vergleichsarbeiten zunächst keine Antwort, sondern es bedarf weitergehender Hilfestellungen und Unterstützungsangebote.

Mit dem kompetenzorientierten Unterricht ist eng verbunden die *Förderung* einzelner Schüler, ausgehend von deren Lernvoraussetzungen. Individuelle Förderhinweise können die Rückmeldungen der Tests aus den bereits genannten Gründen (vgl. Abschnitt 4.3.5) nicht liefern. Dennoch wird zumindest sichtbar, welche Schülergruppen in welchen Bereichen Stärken und Defizite haben. Diese Erkenntnisse lassen sich für die weitere Förderung nutzen, indem zusätzliche Übungsaufgaben eingesetzt und verstärkt binnendifferenzierendes Arbeiten im Unterricht umgesetzt wird.

Als letzten Bereich der Unterrichtsentwicklung ist in diesem Zusammenhang die *Leistungsbewertung* zu nennen. Die Ausrichtung des Unterrichts an den Bildungsstandards verlangt auch ein kompetenzorientiertes Erfassen und Beurteilen der Lernstände der Schüler. Dies erfordert wiederum eine Anpassung der Prüfungsinstrumente. Die Vergleichsarbeiten können in erster Linie hilfreiche Anregungen für Aufgabenstellungen in den schriftlichen Beurteilungsformen, wie den Klassenarbeiten, zur Verfügung stellen (vgl. ebd., S. 97).

Des Weiteren wurde in Abschnitt 5.4.2.1 angeführt, dass über die Nutzung der Tests auch die Evaluationskompetenzen der Lehrkräfte gestärkt werden. So ist es möglich, dass die Lehrenden aus den Rückmeldungen die Konsequenz ableiten, unterrichtsinterne Evaluation intensiver einzusetzen. Auch sind Parallelarbeiten denkbar, mithilfe derer kontinuierlich Vergleiche zwischen den Lernständen der Klassen einer Jahrgangsstufe gezogen und Defizite frühzeitig erkannt werden können (vgl. Schirp, 2006b, S. 433). Für die Umsetzung dieser Maßnahmen ist jedoch die Zusammenarbeit mit dem Kollegium unabdingbar. Generell kann eine Vielzahl der Maßnahmen als Team initiiert und eingeführt werden. Neben gemeinsamen Klassenarbeiten können in diesem Zusammenhang auch die Unterrichtseinheiten in der Form der Kokonstruktion gemeinsam geplant und ausgewertet werden. Dies fördert nicht nur die curriculare Abstimmung der Unterrichtsgegenstände, sondern auch den intensiven Austausch und Konsens über pädagogische, didaktische und methodische Sachverhalte (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 97). Auch Schrader & Helmke (vgl. 2004, S. 156) bestärken diesen Aspekt, da Maßnahmen auf der Grundlage von Ergebnisberichten umso eher realisiert würden, je intensiver die Unterstützung im schulischen Kontext erfolge. Ein kooperatives Vorgehen in der Unterrichtsentwicklung verstärkt sowohl die Motivation als auch die Verbindlichkeit der getroffenen Vereinbarungen.

5.5 Nutzungs- und Wirkungsmodelle

Mittels der Datenrückmeldungen erhalten die Lehrkräfte die Ergebnisse ihrer Lerngruppe aus der Vergleichsarbeit in aufbereiteter Form. Dies allein bewirkt jedoch keinen Anstoß von Schulentwicklung. Es ist vielmehr entscheidend, dass mit den Tests und ihren Rückmeldeberichten intensiv gearbeitet wird. Praktisch bedeutet dies eine Analyse und Interpretation der Ergebnisse, um daran anschließend Handlungsstrategien zu entwerfen und umzusetzen, welche wiederum zu einer Verbesserung der Bildungsqualität an der Einzelschule beitragen sollen. Die Nutzung obliegt daher den Akteuren an der Schule, speziell den betreffenden Lehrpersonen und der Schulleitung. Ausschließlich von der Nutzungsform und -intensität ist es abhängig, in welchem Maße die Vergleichsarbeiten zu einer weiterführenden Schulentwicklung effektiv beitragen können. Hosenfeld führt hierzu an, dass es insbesondere bedeutsam sei, die Verwendung der Tests als einen „Prozess vieler kleiner Schritte zu verstehen, der auch immer wieder Gelegenheiten zur ‚Kurskorrektur' bietet" (2005, S. 114).

Für die theoretische Beschreibung dieses Prozesses dienen Nutzungs- beziehungsweise Wirkungsmodelle, von denen im Folgenden die Modelle von Helmke (2004) und von Visscher und Coe (2003) näher erläutert werden. Als Wirkung werden in diesen Modellen die Effekte der Vergleichsarbeiten auf den verschiedenen Ebenen der Schulentwicklung verstanden. Hierbei können sich direkte Effekte ergeben, indem die Analyse der Ergebnisse unmittelbar in Handlungsstrategien transformiert wird, oder es entstehen indirekte Effekte, so dass bereits bestehende Entwicklungsprozesse verstärkt werden (vgl. van Ackeren, 2003, S. 22). Bei der Wirkung der Vergleichsarbeiten auf die Schulentwicklung ist allerdings zu differenzieren, inwiefern die tatsächlich eintretenden Effekte mit den von den Landesinstituten intendierten Wirkungen übereinstimmen.

Den verschiedenen Nutzungsmodellen ist gemeinsam, dass der Interpretation durch die Lehrenden eine besondere Bedeutung zukommt (vgl. Maier & Rauin, 2006, S. 410). Da den Tests zudem Funktionen auf der Ebene des Systemmonitorings zugeordnet werden, können in diesem Bereich ebenfalls Effekte in Hinblick auf eine Qualitätsverbesserung eintreten. Diese Thematik ist jedoch nicht Gegenstand der Problemstellung dieser Arbeit. Aus diesem Grund werden im Folgenden lediglich Nutzungsmodelle beschreiben, welche sich auf die Wirkung innerhalb der Einzelschule beziehen.

Das zuerst dargestellte Modell „Von der Evaluation zur Innovation - ein Zyklenmodell" von Helmke (2004, S. 100 ff.) beschreibt die einzelnen Nutzungsphasen mit ihren zugrunde liegenden Bedingungen. Im deutschsprachigen Forschungsraum hat sich dieses Modell in besonderer Weise etabliert, so dass es in einem Großteil der bisherigen Forschungsarbeiten

zur Nutzung von Vergleichsarbeiten als theoretische Grundlage herangezogen wurde (zum Beispiel bei Koch, Groß Ophoff, Hosenfeld, & Helmke, 2006; Maier, 2008a; Schneewind, 2007b). Wie in Abbildung 14 zu erkennen ist, besteht das Zyklenmodell, beginnend mit dem Erhalt der Rückmeldung, aus vier Nutzungsphasen, der *Rezeption*, der *Reflexion*, der *Aktion* und der *Evaluation*.

Eine Phase baut hierbei auf der vorangegangenen auf. Dies impliziert zugleich, dass jede Phase als Voraussetzung für die folgende fungiert (vgl. Groß Ophoff, Koch, Hosenfeld, & Helmke, 2006, S. 32). Falls beispielsweise bei der Rezeptions Schwierigkeiten auftreten, die eine weitere Auseinandersetzung mit den Schülerergebnissen verhindern, kann keine Reflexion erfolgen und der Nutzungsprozess bricht an dieser Stelle ab. Um die Komplexität eines solchen Nutzungsprozesses zu verdeutlichen, berücksichtigt das Modell zusätzliche Faktoren, wie die Gestaltung und den Inhalt der Rückmeldung sowie individuelle, schulische und externe Bedingungen, welche auf die Verwendung der Vergleichsarbeiten für die Schulentwicklung in vielfältiger Weise einwirken. Im Folgenden werden die einzelnen Komponenten des Modells erläutert.

Rückmeldung

Die einzelnen Bestandteile eines Rückmeldesystems sowie die Anforderungen an dessen Qualität wurden bereits in Abschnitt 4.3.6.1 hinreichend erläutert. An dieser Stelle ist dennoch zu betonen, dass sowohl die statistische Aufbereitung und Gestaltung als auch die inhaltliche Reichhaltigkeit der Rückmeldung den Nutzungsprozess wesentlich beeinflussen können. Dies geschieht einerseits auf motivationaler Ebene, zum anderen kann das Verständnis der Ergebnisse durch Erläuterungen und Grafiken erleichtert beziehungsweise erschwert werden. Nach Hosenfeld, et al. (vgl. 2006, S. 23) ergäben sich umso mehr potenzielle Nutzungsvarianten, je umfangreicher das Angebot sei. Desto größer sei schließlich auch die Wahrscheinlichkeit, dass die Inhalte mit den individuellen Informationsbedürfnissen der Lehrkraft übereinstimmen würden. Allerdings räumen Hosenfeld, et al. ein, dass mit wachsendem Umfang auch die Gefahr von mangelnder Struktur, Unverständnis und Fehlinterpretation der Informationen steige. Dennoch würde mittels zusätzlicher Daten qualitativer Art, wie didaktischen Hinweisen, die weitere Nutzung der Rückmeldung zusätzlich angeregt werden (vgl. Groß Ophoff, Koch, Hosenfeld, & Helmke, 2006, S. 33).

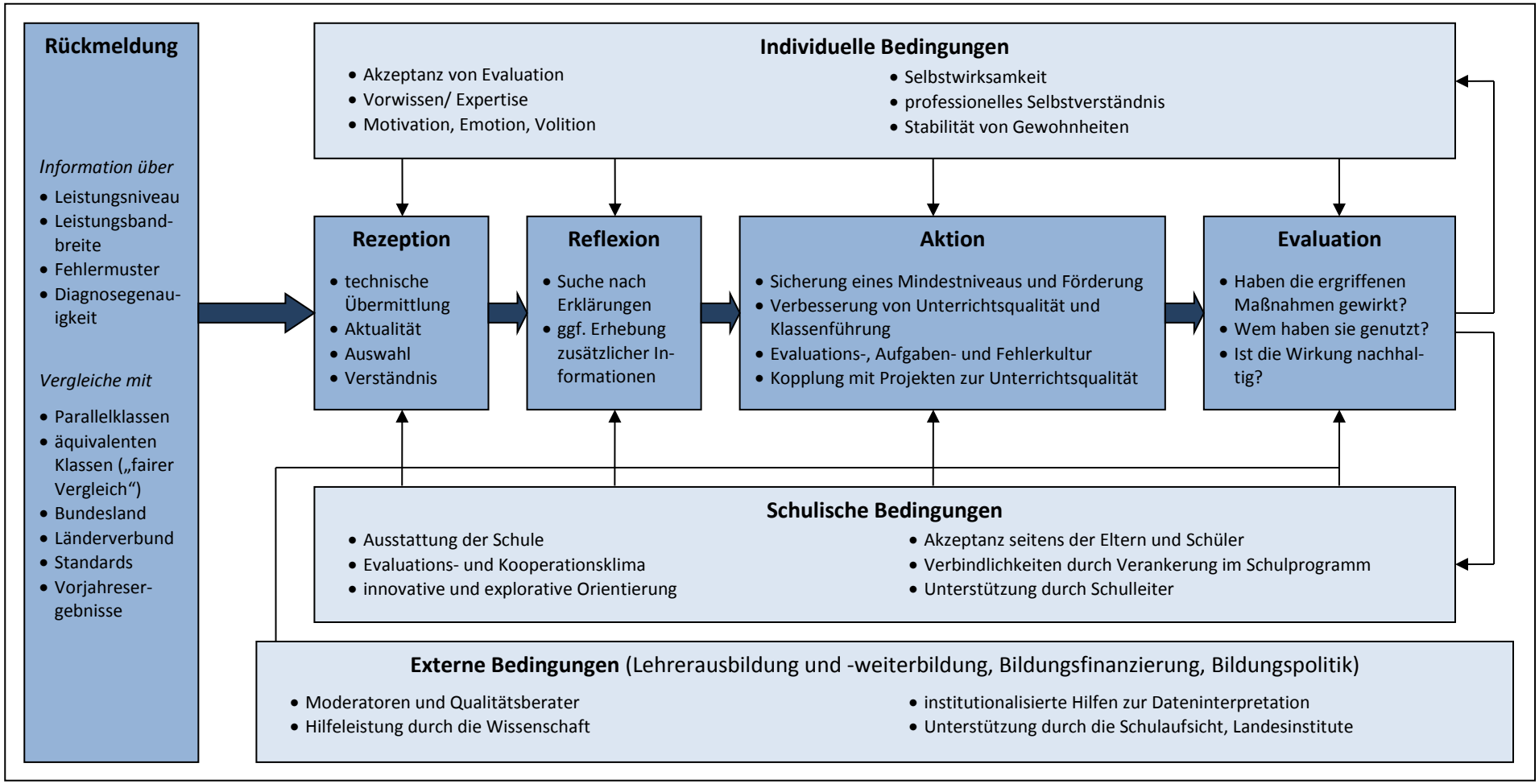

Abbildung 14: Von der Evaluation zur Innovation - ein Zyklenmodell (vgl. Helmke A. , 2004, S. 100)

Rezeption

Die Rezeption umfasst den Erhalt und das Verständnis der Rückmeldungsinhalte sowie die Auswahl der Informationen, welche die individuelle Lehrkraft detaillierter auswerten kann. Für ersteres sind der Übermittlungsweg sowie die Aktualität der Daten zentrale Einflussfaktoren. Für das Verständnis ist hingegen die Akzeptanz der Lehrkräfte gegenüber den Vergleichsarbeiten und den zur Verfügung gestellten Daten entscheidend. Die Motivation, sich mit den Ergebnissen auseinanderzusetzen und sie verstehen zu wollen, kann als Teil des professionellen Selbst betrachtet werden (vgl. Abschnitt 5.4.2.1). Die Aufmerksamkeit erhöht sich zugleich, je mehr Kollegen betroffen sind und sich den Inhalten der Rückmeldung widmen, so dass sich die Rezeptionsbereitschaft auf weitere Personen überträgt. Vorhandene organisationale Strukturen wirken auf diese Weise ebenfalls in die Phase der Rezeption ein. Grundlegende Voraussetzung für das Verständnis ist des Weiteren ein gewisses empirisches Vorwissen sowie eventuell vorhandene Evaluationserfahrungen (Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 24 f.).

Die Art der Rückmeldungsgestaltung kann zudem stark auf das Verständnis einwirken. Falls die Rückmeldung als zu komplex, unerklärlich oder irrelevant eingeschätzt wird, ist ein Abbruch des Nutzungsprozesses die Folge. Für die Akzeptanz ist es jedoch vor allem von Bedeutung, inwiefern die Lehrkräfte die Vergleichsarbeiten mit einem externen Kontrollinstrument und einer ihnen von der Bildungsadministration auferlegten Verpflichtung assoziieren (vgl. Haenisch & Müller, 2005, S. 303; Maier, 2008a, S. 96). Aus diesem Grund muss den Lehrkräften bereits bei der Rezeptionsphase das konkrete Potenzial der Tests für die Schule und den eigenen Unterricht sichtbar sein, so dass sie mit dem Instrument potenzielle positive Wirkungen auf die zukünftige professionelle Arbeit konnotieren (vgl. Kultusministerkonferenz, 2010, S. 14). Hierin liegt zugleich die Herausforderung: Die Vergleichsarbeiten müssen auf eine gewisse Innovationsfreudigkeit im Kollegium stoßen, wobei mit den Tests trotz des Mehraufwandes ein sich lohnender Ertrag verknüpft sein muss. Bonsen, et al. (vgl. 2002, S. 171) wiesen diesbezüglich einen Zusammenhang zwischen der Bereitschaft zu Neuerungen und dem Dienstalter nach, indem Innovationen wie die standardisierten Leistungsmessungen bei Personen mit einer geringeren Anzahl an Dienstjahren eher begrüßt würden.

Reflexion

Anschließend an die Rezeption erfolgt die Reflexionsphase, in der die Schülerergebnisse analysiert und weiterführend die Ursachenkomplexe herausgefiltert werden. Die Reflexion stellt somit eine kommunikative Auseinandersetzung mit möglichen Erklärungsansätzen dar (vgl. Schneewind, 2007b, S. 76). Voraussetzung hierfür ist einerseits das Vorhandensein einer intrinsischen Motivation, die Ergebnisse interpretieren und Handlungen daraus ableiten zu wollen. Andererseits ist für die Intensität des Analyseprozesses auch das professionelle Selbstverständnis der jeweiligen Lehrperson zentral, da die Reflexion eine Offenheit und eine Bereitschaft zu Selbstkritik erfordert (vgl. Tresch, 2007, S. 113). Das Ausmaß der Reflexion wird zudem durch die zur Verfügung stehende Zeit, welche die Lehrkraft für die Auswertung der Testergebnisse aufzubringen bereit ist, beeinflusst (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 29).

Die Reflexion beginnt mit der Identifikation von Stärken und Schwächen der Lerngruppe, indem untersucht wird, welche Ergebnisse besonders auffällig, unerwartet oder erklärungsbedürftig sind. Bei erkennbaren Defiziten sollte zudem betrachtet werden, ob es sich um zufällige oder systematische Fehlerquellen handelt und inwiefern sie sich spezifischen Inhalts- und Kompetenzfeldern zuordnen lassen (vgl. Peek & Dobbelstein, 2006, S. 53). Daran anknüpfend erfolgt eine Bildung von Hypothesen, bei der die ermittelten Schülerleistungen möglichen Ursachenzusammenhängen zugeordnet werden. Hierbei wird zwischen internaler und externaler Attribuierung differenziert.

Bei der internalen Attribuierung werden Ursachen betrachtet, auf welche die Lehrkraft direkt eingewirkt hat, so dass sich die Erklärungen für die Testergebnisse auf die eigene Person beziehen (vgl. Tresch, 2007, S. 124). Hierzu zählen zum einen die Qualität und die Gestaltung des eigenen Unterrichts, wie zum Beispiel die Intensität oder Sequenzierung der geförderten Kompetenzbereiche, der Nutzen von eingesetzten Arbeitsmaterialien, die individuelle Schüler-Lehrer-Beziehung sowie die eigene professionelle Kompetenzentwicklung (vgl. Hosenfeld, 2005, S. 112). Zum anderen werden mittels internaler Attribuierung die persönlichen Erwartungen an die Leistungen der eigenen Lerngruppe reflektiert. Auf diese Weise wird zugleich die diagnostische Kompetenz gefordert und gefördert wird (vgl. Tresch, 2007, S. 113).

Die externale Attribuierung richtet sich hingegen auf die Rahmenbedingungen, auf welche die Lehrperson nur sehr geringen bis keinen Einfluss hat. Die Ursachenzuschreibung kann dabei auf vier verschiedenen Ebenen erfolgen: erstens können auf Schülerebene die Kriterien Vorwissen, Persönlichkeitsmerkmale, individuelle Lernbiografie, Motivation, Anstrengungsbereitschaft, Sprach- und Migrationshintergrund sowie vorhandene Unterstützung

durch das Elternhaus untersucht werden. Auf Klassenebene sind Klassenzusammensetzungen, Unterrichtsausfall und Lehrerwechsel relevante Kontextfaktoren. Des Weiteren ist es denkbar, dass schulbezogene Erklärungen betrachtet werden, wie die verwendeten Lehrwerke, die Abstimmung zwischen Kern- und Schulcurricula oder die Organisation von Förderkonzepten. Zuletzt können auch externe Ursachenerklärungen einfließen, wie beispielsweise die Konstruktion der Vergleichsarbeiten, die Instruktionsform bei der Durchführung sowie der Schwierigkeitsgrad der Items (vgl. Ballasch, 2009, S. 302 f.; Hosenfeld, 2005, S. 112; Peek & Dobbelstein, 2006, S. 52 f.).

Die in der Reflexion aufgestellten Hypothesen bedürfen einer sich anschließenden Prüfung. Bereits vorhandene zusätzliche Informationen wie vergangene Klassenarbeiten, Zeugnisnoten oder Parallelarbeiten können in diesem Zusammenhang weitere Erkenntnisse liefern. Gegebenenfalls bietet sich zudem die Durchführung von spezifischen unterrichtsinternen Evaluationen an. Generell sollte die Reflexionsphase in einem möglichst kooperativen und kommunikativen Arbeitsprozess verlaufen. Über einen Abgleich der Fehlermuster über die Lerngruppe hinweg mit denen der Parallelklassen und die gemeinsame Diskussion dieser Auffälligkeiten können weitergreifende Erklärungen für die Schülerleistungen gefunden werden (vgl. Peek & Dobbelstein, 2006, S. 52 f.).

Tresch (vgl. 2007, S. 123) führt bezüglich der Intensität des Reflexionsprozesses an, dass vor allem dann ausführliche Analysen vorgenommen würden, wenn die Ergebnisse erwartungswidrig sind oder eine Diskrepanz zu der eigenen Leistungseinschätzung der Lerngruppe aufzeigen. In der Attribuierungstheorie wird des Weiteren davon ausgegangen, dass positive Testergebnisse tendenziell eher internal attribuiert werden, während die Lehrpersonen bei defizitären Ergebnissen zu externaler Attribuierung neigen (vgl. ebd., S. 124). Diese Erscheinung kann mit einem geringer entwickelten professionellen Selbstverständnis zusammenhängen: wenn sich die Lehrpersonen bei schlechteren Leistungen in der Ausübung ihrer Arbeit angegriffen fühlen, schreiben sie als Form des Selbstschutzes die Gründe eher externalen Bedingungen zu (vgl. Schneewind, 2007b, S. 61). Dies entlastet einerseits, da die Verantwortung für die Ergebnisse abgegeben wird. Andererseits erschwert dies den weiteren Nutzungsprozess der Vergleichsarbeiten erheblich, denn aufgrund fehlender Selbstreflexion können keine Handlungsstrategien entworfen werden und es kann weiterführend keine Qualitätsentwicklung stattfinden (vgl. Tresch, 2007, S. 125). Aus diesem Grund lässt sich schlussfolgern, dass defizitäre Ergebnisse zwar zu einer verstärkten Reflexion anregen können, aber im Fall einer überwiegend externalen Attribuierung die Entwicklung von Handlungskonzepten nicht zwangsläufig erfolgen muss beziehungsweise die Handlungsbereitschaft bereits beeinträchtigt ist. Zudem ergeben sich bei einer externalen Ursa-

chenzuschreibung für die Lehrenden Konfliktfelder, denn sie übernehmen die Verantwortung für die Ergebnisse nicht selbst, werden aber von der außerschulischen Öffentlichkeit per sé als die für die Leistungen der Schüler verantwortlichen Personen betrachtet (vgl. Schneewind, 2007b, S. 59).

Dies impliziert zugleich Anforderungen an die Rückmeldung. Einerseits sollten sie Informationen beinhalten, die der jeweiligen Lehrkraft bislang unbekannt sind und die eine weiterführende Reflexion anregen, denn letztlich stehen die Sinnhaftigkeit und der Verwertungseffekt von Rückmeldungen in starker Abhängigkeit zu dem Erkenntnisgewinn für die Lehrperson (vgl. ebd., S. 83). Andererseits sollten die Rückmeldungen eine interne Attribuierung unterstützen, indem sie als Möglichkeit betrachtet werden, das unterrichtliche Agieren und die eigene Leistung als Lehrperson selbst beurteilen zu können (vgl. Maier, 2008a, S. 98).

Aktion

Die Reflexion selbst führt noch nicht zu einer Qualitätsverbesserung von Schule und Unterricht. Hierfür ist die Phase der Aktion zwingend erforderlich, in der die Handlungsstrategien entworfen und umgesetzt werden. Die aus den Rückmeldungen abgeleiteten Konsequenzen können sehr vielfältig sein und sowohl in den Bereich der Unterrichtsentwicklung als auch in die Organisations- und Personalentwicklung einwirken. Daher ist eine Auflistung aller denkbaren Handlungskonzepte an dieser Stelle nicht möglich. Eine Benennung der Aktionen, wie sie beispielsweise in den Handreichungen für die Lehrkräfte von den Landesinstituten dargelegt werden, ist den Ausführungen von Ballasch (vgl. 2009, S. 303) und dem Hessischen Kultusministerium (vgl. 2009, S. 24) zu entnehmen.

Jede Handlung sollte letztlich der Qualitätssicherung und -entwicklung an der Schule dienen. Die Aktionsphase umfasst generell mehrere Arbeitsschritte. Zunächst müssen aus der Ergebnisanalyse Zielsetzungen formuliert werden, anhand derer wiederum geeignete kurz-, mittel- und langfristige Maßnahmen herausgefiltert werden. Je kleinschrittiger und konkreter die Ziele benannt werden, desto passgenauer können die hierfür notwendigen Handlungen identifiziert werden. In Hinblick auf eine nachhaltige Schulentwicklung ist es zielführend, wenn die Maßnahmen mit bereits bestehenden Qualitätsprozessen vernetzt oder in weitere Evaluationsprojekte, wie die Schulinspektion, integriert werden, so dass Synergieeffekte entstehen können. Die Implementierung der Maßnahmen bedarf zudem einer spezifischen Planung in Hinblick auf die notwendigen Ressourcen und den zeitlichen Umfang (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 30 ff.). Für den Erfolg des Nutzungsprozesses ist letztlich die Umsetzung der initiierten Handlungen von entscheidender Bedeutung. Bei gelingender Rezeption und Reflexion können Erkenntnisse abgeleitet und die Motivation zu

Veränderungen entwickelt worden sein. Jedoch muss diese Handlungsbereitschaft auch zwingend in die Tat umgesetzt werden. Erst mit der Aktion kann der Nutzen aus der Rezeption und der Reflexion sichtbar und qualitative Entwicklungsprozesse angestoßen werden. Schneewind bezeichnet die Aktion daher auch als das „Kernstück des Prozesses“ (vgl. 2007b, S. 81).

Evaluation

Gleichermaßen, wie die Evaluation einen obligatorischen Bestandteil aller initiierten Maßnahmen der Schulentwicklung darstellen sollte, ist sie auch für die Nutzung und Wirkung der Vergleichsarbeiten als abschließender Schritt bedeutsam. Mittels geeigneter Methoden sollte überprüft werden, inwiefern die eingeleiteten Handlungen wirken und die intendierten Zielsetzungen erreicht worden sind. Die Erfahrungen sowie die Relevanz der Evaluationsergebnisse für die weitere Schulentwicklung sind ebenfalls zu betrachten (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 32).

Bedingungen

Der Nutzungs- und Wirkungsprozess der Vergleichsarbeiten wird nach dem Modell von Helmke von individuellen, schulischen und externen Rahmenbedingungen zusätzlich beeinflusst.

Als eine zentrale individuelle Bedingung ist der Entwicklungsgrad des professionellen Selbst anzuführen. Die einzelnen Komponenten dieses Konstrukts wurden bereits im Abschnitt 5.4.2.1 ausführlich erläutert und konnten teilweise auch der Abbildung 14 entnommen werden. Besonders wichtig ist in diesem Zusammenhang die Bereitschaft zu (Selbst-) Reflexion sowie das Empfinden der professionellen Verpflichtung, für die Leistungen der Lerngruppe Verantwortung zu übernehmen und gegebenenfalls bei defizitären Ergebnissen unverzüglich Maßnahmen zur Steigerung der Lernqualität und -effektivität zu ergreifen. Zudem sind Fähig- und Fertigkeiten, wie bereits vorhandene Evaluationskompetenzen in Bezug auf die Dateninterpretation oder Erfahrungen mit differenzierenden Unterrichtsansätzen, von Bedeutung. Nicht zu unterschätzen ist des Weiteren die individuelle innere Einstellung gegenüber den Vergleichsarbeiten als ein externes Messinstrument. Ist die Lehrperson dieser Innovation gegenüber aufgeschlossen, so erleichtert dies die einzelnen Nutzungsphasen. Stößt der Test jedoch auf mangelnde Akzeptanz oder auf eine demotivierende Erwartungshaltung, wird die produktive Arbeit mit den Ergebnissen erheblich erschwert (vgl. Posch, 2009, S. 130 f.).

Zu den schulischen Bedingungen zählen bereits vorhandene organisationale Strukturen, wie die Etablierung von Fachgremien oder das generelle Professionswissen an der Schule (vgl. Hartung-Beck, 2009, S. 33). Einfluss hat auch die Einführung vorangegangener Evaluationsinstrumente und deren Akzeptanz sowohl im Kollegium als auch in der Eltern- und Schülerschaft. Wie bereits in Abschnitt 5.4.1.2 erläutert wurde, ergeben sich auch aus der Innovationsoffenheit der Schulleitung gegenüber outputorientierten Reformmaßnahmen sowie aus der in kooperativen Strukturen verankerten Verbindlichkeit der Tests Auswirkungen auf das Ausmaß und die Intensität des Nutzungsprozesses.

Als externe Bedingungen sind insbesondere die Unterstützungssysteme der Lehreraus- und fortbildung sowie die von der Bildungspolitik zur Verfügung gestellten finanziellen und materiellen Ressourcen anzuführen (vgl. Hosenfeld, Groß Ophoff, & Bittins, 2006, S. 35). Hierzu zählen beispielsweise die Handreichungen für die Durchführung, Auswertung und Verwendung der Tests sowie didaktische Erläuterungen, in denen die Vergleichsarbeiten für die Lehrkräfte transparent in den Reformkontext eingebettet werden. Denkbar sind auch Datenbanken mit Testaufgaben oder direkte Beratungsangebote während der Auswertungsphase der Rückmeldungen. Diese Maßnahmen sollen den Lehrpersonen den Umgang mit den Ergebnissen nicht nur erleichtern und die Selbstreflexion befördern, sondern auch zur Entwicklung einer inneren Verpflichtungshaltung zu kontinuierlicher Evaluation und Schulentwicklung beitragen (vgl. Posch, 2009, S. 126). Insbesondere mittels Fortbildungen können die Landesinstitute gezielt die Akzeptanz der Lehrerschaft gegenüber den Tests steuern, indem Evaluationskompetenzen entwickelt und die Lehrkräfte dazu befähigt werden, die Daten spezifisch interpretieren zu können (vgl. Fussangel, Rürup, & Gräsel, 2010, S. 330 f.). Diese Unterstützung kann sowohl durch administrative Stellen, wie den Landesinstitute, dem Schulamt und der Schulaufsicht, als auch durch externe Dienstleister, Beratungszirkel, Netzwerke und wissenschaftliche Experten erfolgen. Allerdings kann die angebotene Unterstützung auch als eine Form der Kontrolle von den schulischen Akteuren wahrgenommen werden, was infolgedessen die Akzeptanz des Testinstruments massiv negativ beeinflussen würde. Die Aufgaben der Landesinstitute entsprechen hingegen eher dem Informationsmanagement und dem Angebot von Beratung und Hilfestellung im Bedarfsfall (vgl. Klieme, et al., 2007, S. 114).

Resümierend ist somit festzuhalten, dass bisherige Entwicklungen der Schulqualität in vielfältiger Hinsicht den Nutzungsprozess der Vergleichsarbeiten sowohl befördern als auch hemmen können. So ist beispielsweise für eine kollegiale Reflexion die bereits vorhandene Etablierung kooperativer Strukturen in der Organisation grundlegend. Demgegenüber wur-

de ausführlich dargelegt, dass die Vergleichsarbeiten wiederum das Ziel verfolgen, bereits bestehende Qualitätsprozesse zu katalysieren bzw. neue Entwicklungen zu forcieren. So sind Schulentwicklungsprozesse sowohl Voraussetzung als auch Zielgegenstand bei der Nutzung der Tests.

Kritisch sollte an dieser Stelle eingeräumt werden, dass der Organisationsentwicklung besondere Bedeutung zugemessen wird, da die Fachgruppen und -konferenzen neben den Lehrkräften als primäre Adressaten der Vergleichsarbeiten benannt werden. Hiermit ist die Intention verbunden, dass mittels verbesserter organisationaler Strukturen und steigender Professionalisierung auch eine Qualitätsverbesserung des Unterrichts einhergehe. Den outputorientierten Reformen liegt als Zielsetzung die Forderung nach der Verbesserung der Schülerleistungen in der Bundesrepublik zugrunde. Daher sollten auch die Vergleichsarbeiten primär die Förderung der Unterrichtsentwicklung forcieren und den Lehrpersonen in diesem Bereich wertvolle, unterstützende Informationen liefern. Der Anstoß von Organisations- und Personalentwicklung stellt aus diesem Grund lediglich ein indirektes Zwischenziel dar, dessen tatsächlicher Effekt auf die Unterrichtsentwicklung unklar bleibt.

Der Wert des Zyklenmodells nach Helmke „liegt in seiner heuristischen Kraft, die verdeutlicht, dass eine Rückmeldung alleine keineswegs ausreicht, sondern spezifische Bedingungen des Gelingens erfüllt sein müssen um (insbesondere auch kontinuierlich) Veränderungen zu bewirken“ (Hosenfeld & Groß Ophoff, 2007, S. 358). Mittels des Modells wird die Komplexität von hypothetisch zu erwartenden beziehungsweise empirisch bereits nachgewiesenen Einflussfaktoren auf den Wirkungsprozess von standardisierten Tests wie den Vergleichsarbeiten offenbar (vgl. Maier, 2008a, S. 96). Die Ableitung von Handlungskonsequenzen vollzieht sich nicht automatisch, sondern die „Prozesse der […] Schulentwicklung können nur durch eine Kombination aus validen und an die Handlungsebene anschlussfähigen Informationen, Unterstützungssystemen sowie Kooperations- und Änderungsbereitschaft, zeitlichen Ressourcen, gut überlegten und langfristig strukturierten Maßnahmen durch das Team des Schulhauses insgesamt in Gang gesetzt werden“ (Schneewind, 2007b, S. 48). Dagegen umfasst die folgende Auflistung resümierend die zentralen Hemmnisse bei der Nutzung (vgl. Kohler, 2005, S. 56 ff.):

- fehlende Akzeptanz aufgrund einer objektiven sowie subjektiv empfundenen erhöhten Arbeitsbelastung,
- fehlende Unterstützungssysteme, welche die Lehrer zu einer professionellen Nutzung anleiten,
- gering ausgeprägte kommunikative und kooperative Strukturen,

- unzureichende Selbstwirksamkeitserwartungen, welche zu einer Handlungspassivität führen, sowie
- Voreingenommenheit in der Attribuierung.

Als weiteres Wirkungsmodell soll an dieser Stelle das Rahmenmodell wichtiger Einflussfaktoren auf School Performance Feedback Systems von Visscher & Coe (vgl. 2003, S. 331) vorgestellt werden. Die zentralen Elemente des Rahmenmodells sind in Abbildung 15 ersichtlich.

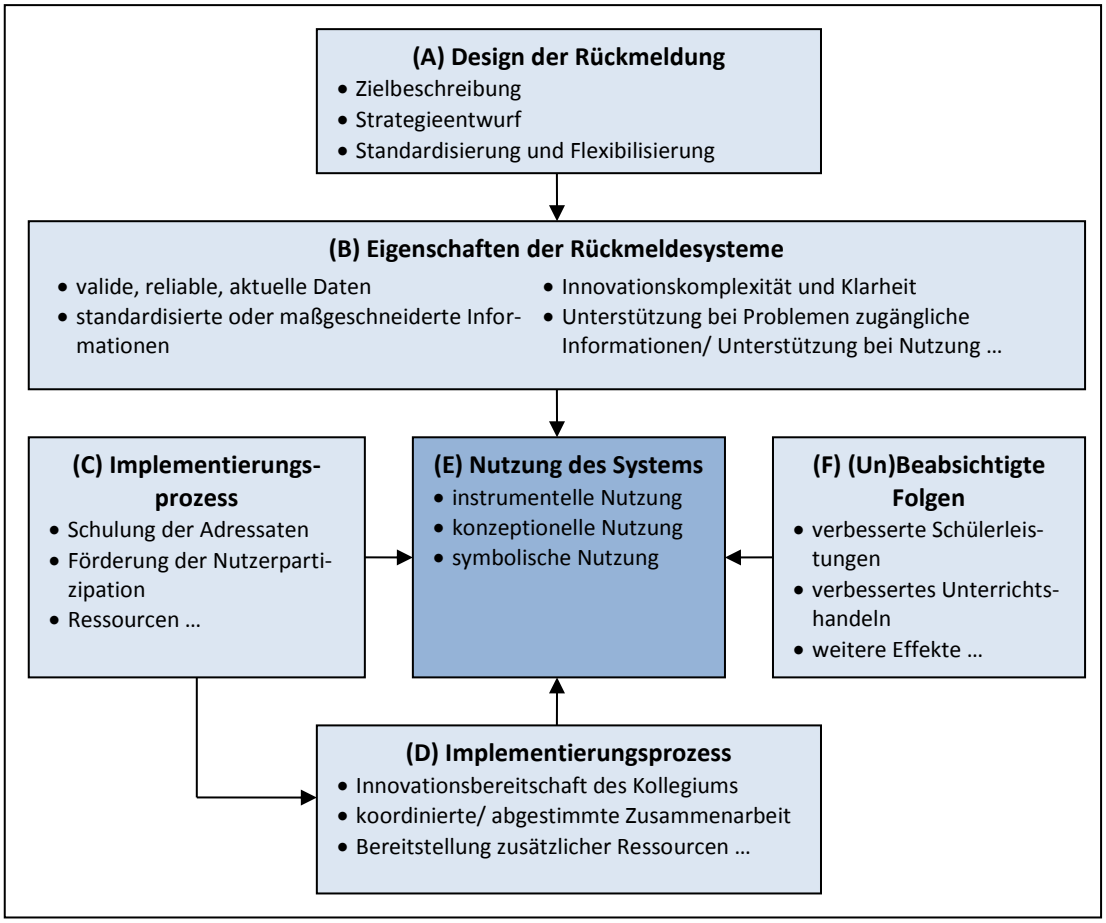

Abbildung 15: Rahmenmodell wichtiger Einflussfaktoren auf School Performance Feedback Systems (vgl. Visscher & Coe, 2003, S. 331; Bonsen & von der Gathen, 2004, S. 249)

Dieses Modell ist nicht aus der Perspektive der einzelnen Lehrkraft, sondern aus einem gesamtschulischen Blickwinkel heraus konstruiert. Äquivalent zu dem Modell von Helmke wird auch hier der Einfluss der Rückmeldungsgestaltung thematisiert. Als zentrale Rahmenbedingungen für die Nutzung werden allerdings weniger die individuellen Voraussetzungen betrachtet, sondern vielmehr schulische Kontextfaktoren sowie administrative Vorgaben, mit denen die Schulen durch die Implementierung der Tests konfrontiert werden. Bezüglich

der Bedingungen und eintretenden Wirkungen beschränkt sich das Modell hingegen ausschließlich auf die Unterrichtsentwicklung. Aus diesem Grund ist den Ausführungen von Altrichter (vgl. 2010, S. 231) zuzustimmen, dass das Rahmenmodell den Mehrebenencharakter der Schulorganisation und -entwicklung sowie die Handlungsmöglichkeiten einzelner Personen nur ungenügend berücksichtige.

Für den Erkenntnisgewinn dieser Arbeit sind allerdings die verschiedenen Nutzungsformen von Interesse, welche im Zentrum des Modells stehen und auf Rossi, et al. (vgl. 2004, S. 411 f.) zurückzuführen sind. Bei der *instrumentellen Nutzung* werden aufgrund der Informationen, die aus der Rückmeldung entnommen werden können, direkte Entscheidungen getroffen und Handlungen abgeleitet. Die Ergebnisse von Reflexion und Rezeption werden hierbei in Taten sichtbar, während bei der *konzeptionellen Nutzung* lediglich das Denken oder die inneren Einstellungen der jeweiligen Person beeinflusst werden. Die Wirkungen aus den Tests sind somit unbewusste Effekte. Bei der *symbolischen Nutzung* werden die Rückmeldungen ausschließlich selektiv dazu verwendet, die eigenen Einschätzungen und Beurteilungen, welche sich bereits vor dem Test manifestiert hatten, argumentativ zu bestätigen. Dies führt jedoch nicht zu einer Handlungsbereitschaft (vgl. Maier & Rauin, 2006, S. 406).

5.6 Typisierung der Nutzungsformen

Aufgrund der Kategorisierung in diverse Nutzungsformen lassen sich die möglichen Reaktionsweisen der jeweiligen Lehrkraft bei der Verwendung der Vergleichsarbeiten und ihrer Rückmeldungen herauskristallisieren. Schneewind (vgl. 2007b, S. 45 f.) ermittelte hierbei sieben verschiedene Verhaltensmuster, die im Folgenden kurz dargelegt und mit den Nutzungsformen nach Visscher & Coe (2003) sowie mit den Nutzungsphasen nach Helmke (2004) verknüpft werden:

1. Die Lehrkräfte verstehen die Rückmeldung nicht und legen sie beiseite. Der Rezeptionsprozess ist nicht gelungen und die weitere Nutzung bricht ab.

2. Die Lehrkräfte nehmen die Informationen zur Kenntnis, leiten hieraus jedoch keine Maßnahmen ab. Der Nutzungsprozess endet mit der Reflexion; eine konzeptionelle Nutzung ist dennoch möglich.

3. Die Lehrkräfte verstehen die Informationen, verwenden sie im Sinne einer symbolischen Nutzung ausschließlich zur Bekräftigung ihrer Voreinschätzungen und leiten keine weiterführenden Handlungen ab.

4. Die Lehrkräfte verstehen die Informationen und vergleichen sie analysierend mit den eigenen Beurteilungen. Es liegt somit zunächst eine konzeptionelle Nutzung vor, bei welcher die diagnostische Kompetenz gefördert wird. Die Phase der Aktion muss jedoch nicht zwingend erfolgen.

5. Die Lehrkräfte verstehen die Informationen und nehmen sie als einen Handlungsimpuls wahr, so dass sie Maßnahmen hieraus ableiten und umsetzen. Dies entspricht der instrumentellen Nutzung und der Aktionsphase.

6. Die Lehrkräfte stellen die Informationen infrage und überprüfen deren Richtigkeit. Bei negativem Ergebnis sinkt die Akzeptanz für das Testinstrument und die weitere Nutzung bricht ab. Dennoch liegt hier eine konzeptionelle Nutzung vor, indem Evaluationskompetenzen geschult werden.

7. Die Lehrkräfte lehnen die Informationen generell ab. Eine Nutzung findet in keiner Weise statt.

Aus diesen verschiedenen Reaktionsmustern wird ersichtlich, dass ausschließlich bei der fünften Option eine instrumentelle Nutzung stattfindet und der Arbeitsprozess die Phase der Aktion erreicht. Bei allen anderen Verhaltensweisen bricht der Nutzungsprozess bereits zuvor ab. Dies hat enorme Konsequenzen auf das Potenzial der Vergleichsarbeiten, auf die weitere Schulentwicklung systematisch und strukturiert einwirken zu können.
Von Relevanz ist in diesem Zusammenhang die qualitative Interviewstudie von Hartung-Beck (2009), welche die Zielsetzung verfolgte, für die Nutzung der Vergleichsarbeiten - speziell für die Lernstandserhebungen in Nordrhein-Westfalen - spezifische Professions- und Organisationstypen herauszufiltern. Die Studie beschränkt sich somit auf die Bereiche Personal- und Organisationsentwicklung. Im Zentrum stand die Fragestellung, in welcher Form vorhandene Strukturen und professionelle Überzeugungen die Nutzungsform und somit auch die Wirkungen der Tests beeinflussen können. Hartung-Beck ermittelte hierbei anhand diverser Merkmalsräume jeweils vier verschiedene Professions- und Organisationstypen, deren charakteristische Merkmale kurz dargelegt werden (vgl. ebd., S. 125 ff.):

Professionstypen

1. Typ A – technologisch, operational

   Dieser Typus versteht die Vergleichsarbeiten als ein objektives Diagnoseinstrument und als einen wertvollen Bestandteil der unterrichtlichen Arbeit. Es wird eine ausführliche Ursachenanalyse vorgenommen, bei der die Erklärungen überwiegend internal der eigenen Person zugeschrieben werden. Konsequenzen für das weitere professionelle Handeln werden abgeleitet.

2. Typ B – technologisch, positionell

   Für den Typ B ist bei der Analyse insbesondere der soziale Vergleich mit Referenzwerten von Bedeutung, wodurch die Rezeption und Reflexion weniger intensiv erfolgt. Die Ursachen für die Ergebnisse werden verstärkt external attribuiert. Dennoch werden viele Handlungsstrategien entworfen, die jedoch kaum nachhaltige Wirkungen erzielen.

3. *Typ C – normativ, operational*

   Typ C befürwortet die neuartige Konzentration auf die Vermittlung von Kompetenzen, empfindet die Reformkonzepte jedoch teilweise als nicht der Schulwirklichkeit entsprechend. Rezeption und Reflexion finden vor allem bei erwartungswidrigen Ergebnissen statt, wobei die Ursachenanalyse nicht eindeutig ist und internale Attribuierungen vermieden werden, da die Lehrkraft den Informationen nur bedingt vertraut. Dennoch werden mittelbare Handlungen abgeleitet. Kritisch wird die zeitliche Belastung betrachtet, die mit der Auswertung verbunden ist.

4. *Typ D - normativ, positionell*

   Einer externen Evaluation wird von Typ D statt Vertrauen eher Skepsis entgegengebracht. Die Vergleichsarbeiten werden daher nicht als unterstützendes Diagnoseinstrument wahrgenommen, sondern als Beurteilung und Kontrolle. Folglich schreibt Typ D die Ursachen für die Ergebnisse eher schwer beeinflussbaren Rahmenbedingungen auf Schülerebene zu. Ein Zusammenhang mit der Qualität der eigenen Arbeit wird nicht hergestellt, so dass auch keine eindeutigen Maßnahmen ergriffen werden.

Aus der Betrachtung der vier Professionstypen kann geschlussfolgert werden, dass einzig Typ A den Vorstellungen entspricht, welche die Landesinstitute mit der Nutzung der Vergleichsarbeiten verknüpfen, währenddessen alle anderen Typen Möglichkeiten jenseits der

ursprünglichen Absichten ergreifen. Das Handeln des Typs D weist hierbei die stärkste Abweichung von den Intentionen zur Testverwendung auf (vgl. ebd., S. 143).

Organisationstypen

1. Typ 1 – autonom, kollektiv

   Typ 1 versteht die Vergleichsarbeiten als ein Instrument zur eigenen autonomen Kontrolle und zur Weiterentwicklung des professionellen Selbst, so dass eine intensive Selbstreflexion stattfindet. Zusätzlich erfolgt die Nutzung in kooperativer Art und Weise, aus welcher vielfältige Handlungskonzepte abgeleitet werden.

2. Typ 2 – autonom, individuell

   Die Ergebnisse der Vergleichsarbeiten werden von Typ 2 ebenfalls eigenständig mit der Qualität seiner individuellen Arbeit verknüpft, jedoch beschränken sich die initiierten Maßnahmen auf den eigenen Unterricht. Kooperation findet nur in Form formaler Absprachen statt.

3. Typ 3 – heteronom, kollektiv

   Typ 3 betrachtet die outputorientierten Reformmaßnahmen als bürokratisch und von außen aufoktroyiert. Den Sinngehalt und das Potenzial der Vergleichsarbeiten stellt er infrage. Dennoch setzt sich Typ 3 schematisch und kollektiv mit den Ergebnissen auseinander. Die Nutzung wird jedoch als Einengung des eigenen Entscheidungsraumes wahrgenommen und die damit verbundene zeitliche Belastung als sehr hoch eingeschätzt.

4. Typ 4 – heteronom, individuell

   Typ 4 empfindet die Vergleichsarbeiten als eine fremdbestimmte Kontrolle und führt nur die notwendigsten Schritte aus. Eigenständige Konsequenzen werden nicht aus den Ergebnissen abgeleitet. Auch wird die Zusammenarbeit nicht gewünscht, so dass kollektive Absprachen vermieden werden (vgl. ebd., S. 147 ff.).

Gleichsam wie bei den Professionstypen entspricht der zuerst vorgestellte Typus 1 am ehesten den Intentionen der Nutzungsform von Vergleichsarbeiten, währenddessen bei Typ 4 der stärkste Widerstand gegenüber den Tests vorherrscht und eine nachhaltige Nutzung daher sehr unwahrscheinlich ist (vgl. ebd., S. 163). Hartung-Beck ermittelte zusätzlich die möglichen Kombinationsmöglichkeiten der dargelegten Professions- und Organisationstypen. Folglich stellt der Typ A/1 die Idealform bezüglich der Verwendung der Tests dar, während die Wirkungen bei D/4 vermutlich eher gering sind (vgl. ebd., S. 160 ff.). Zu erwähnen

ist in diesem Zusammenhang, dass in der durchgeführten Interviewstudie nicht alle kombinatorisch möglichen Professions- und Organisationstypen angetroffen wurden (vgl. ebd., S. 164).

Darüber hinaus kann mit Hilfe dieser Typisierung das Argument verstärkt werden, dass die Nutzung der Vergleichsarbeiten und somit auch deren Potenzial für die Schulentwicklung in einem hohen Maße von dem professionellen Selbstverständnis der betroffenen Lehrkräfte sowie von den bereits vorhandenen organisationalen Strukturen abhängen. Aus den Ausführungen wird zudem deutlich, dass jeweils nur ein Nutzungstyp beziehungsweise ein spezifisches Reaktionsmuster zu einer direkten Verwendung der Tests führt, die wiederum in die Umsetzung innovativer Handlungsstrategien mündet. Folglich stellt sich natürlich die Frage, wie groß der Anteil dieses Professionstyps in der Lehrerschaft ist. Es kann vermutet werden, dass bei einem Großteil der betreffenden Lehrpersonen die Vergleichsarbeiten nicht in dem intendierten Sinne verwendet werden. Interessant ist daher, die Ergebnisse bisheriger Forschungsuntersuchungen zu der Nutzung standardisierter Tests zu betrachten.

5.7 Bisherige Forschungsergebnisse zur Nutzung standardisierter Schulleistungsmessungen für die Schulentwicklung

Die Intentionen der Landesinstitute zu der Nutzung der Vergleichsarbeiten durch die schulischen Akteure stellen gewissermaßen einen Idealtypus dar, den es anzustreben gilt. Es ist im besonderen Maße von Bedeutung zu untersuchen, wie und in welcher Form die Lehrkräfte und die Schulleitungen in der Realität mit den Leistungstests agieren, um daraus schlussfolgern zu können, inwiefern das Instrument seinen Funktionen für die Schulentwicklung tatsächlich gerecht wird. Die Rezeptionsforschung befasst sich daher mit den Bedingungen und den Verläufen der Nutzungsprozesse. Je nach Zielstellung der Untersuchungen werden Einstellungen, Akzeptanz und entwickelte Handlungsstrategien erfasst.

Bislang gibt es für die in der Bundesrepublik Deutschland durchgeführten Vergleichsarbeiten nur vier Rezeptionsstudien, so dass die vorhandenen Informationen bezüglich der wirklichen Nutzung der Rückmeldungen durch die Schulakteure sehr begrenzt sind. Hinzu kommt, dass ein direkter Vergleich der Ergebnisse dieser Untersuchungen schwer möglich ist, da die Studien teilweise zu einem Zeitpunkt vorgenommen wurden, an dem die Testentwicklung noch nicht zentral durch das IQB erfolgte. Aus diesem Grund können die Zielsetzungen und Gestaltungen der betrachteten Tests voneinander abweichen. Ebenfalls ist zu berücksichtigen, dass die Rückmeldungen, welche einen zentralen Einfluss auf die Akzeptanz und die Nutzungsbereitschaft ausüben (vgl. Abschnitte 4.3.6.1 und 5.5), in jedem Bun-

desland in ihrer inhaltlichen Konzeption und Gestaltung variieren. Die Grundvoraussetzungen für die Verwendung der Ergebnisse sind daher bei den verschiedenen Studien nicht identisch. Obwohl die systematische Erforschung der Wirkung der Vergleichsarbeiten erst beginnt, liefern diese Untersuchungen bereits erste bedeutsame Informationen, ob und in welcher Weise die Vergleichsarbeiten bislang verwendet wurden.

Zudem existieren Forschungsergebnisse zur Rezeption von weiteren standardisierten Testverfahren, welche entweder der Zielsetzung des Bildungsmonitorings oder der Individualdiagnostik zuzuordnen sind. Aufgrund dieser funktionellen Differenzen ist eine Übertragung der Ergebnisse auf die Nutzung der Vergleichsarbeiten erschwert. Des Weiteren differieren Testdesign, Stichprobenumfang und methodische Durchführung der verschiedenen Leistungsmessungen teilweise fundamental voneinander (vgl. Maier, 2008a, S. 99; Maier, 2010a, S. 113). Insbesondere die Rezeptionsstudien zu den Monitoring-Tests können in Hinblick auf die Nutzung für die individuelle Schulentwicklung nur geringfügig Informationen liefern, denn die Einzelschule steht nicht im Zentrum deren Zielsetzung (vgl. Hartung-Beck, 2009, S. 34). Da die Erkenntnisse zur Rezeption der Vergleichsarbeiten bislang begrenzt sind, empfiehlt sich dennoch eine Betrachtung der zentralen Ergebnisse weiterer ausgewählter Forschungsuntersuchungen zur Nutzung standardisierter Tests im deutschsprachigen Raum. Rezeptionsstudien aus den USA oder England werden hierbei unberücksichtigt gelassen, da die Konzeption der dortigen Tests zu stark von derjenigen der Vergleichsarbeiten abweicht. Eine Zusammenfassung der wesentlichen angloamerikanischen Forschungsergebnisse kann bei Maier (2010a) nachgelesen werden.

5.7.1 Rezeptionsergebnisse zu Bildungsmonitoring-Tests

TIMSS

Bei der TIMSS-Studie erhielten die Lehrpersonen eine Rückmeldung, welche die Klassenmittelwerte in Beziehung zu den nationalen Durchschnittsreferenzwerten setzte. Anschließend wurden Lehrkräfte, Eltern und Schulaufsicht mittels eines Fragebogens zu Einstellungen und der Reflexion der Ergebnisse befragt. Es war zwar eine hohe Bereitschaft vorhanden, an externen Leistungsstudien teilzunehmen, jedoch bewerteten die Lehrpersonen die Ergebnisse nur als mäßig bedeutsam für ihre eigene Unterrichtsarbeit. Bei allen drei Personengruppen wurden interessanterweise externale Attribuierungsvoreingenommenheit festgestellt, was sich negativ auf die Intensität der Reflexions- und Aktionsphase auswirkte (vgl. Kohler, 2009, S. 85 ff.).

PISA-2000

Im Zuge der PISA-2000-Studie konnten die Schulen eine freiwillige Rückmeldung in Form von Broschüren mit schulspezifischen Informationen erhalten. Zur Rezeption befragte das Projektbüro Vergleichsuntersuchungen des Hessischen Landesinstituts für Pädagogik schriftlich 106 Schulen in Hessen, die an PISA teilgenommen hatten. Der Rücklauf betrug 68 Prozent. Bezüglich der Nutzung von Unterstützungssystemen konnte ermittelt werden, dass Schulen, die eine zusätzliche Beratung in Anspruch genommen hatten, größere Planungsperspektiven und eine verstärkte Handlungsbereitschaft für die Auseinandersetzung mit den Ergebnissen aufwiesen. Der Schulberatung wurden positive Verstärkungstendenzen auf den Nutzungsprozess zugeschrieben (vgl. Markstahler, Schwarz, & Steffens, 2004, S. 200 ff.).

IGLU

Die freiwillige IGLU-Rückmeldung beinhaltete Klassenwerte im Vergleich zu nationalen Daten. Bei einer Fragebogenuntersuchung mit einem Rücklauf von 20 Prozent und einer Beteiligung von 53 Lehrpersonen konnte eine positive Beurteilung des Rückmeldesystems konnotiert werden. Demnach traten keine bis wenige Verständnisschwierigkeiten bei der Rezeption auf. Die Lehrkräfte äußerten dennoch den Wunsch nach einer weiterführenden Beratung (vgl. Schwippert, 2004, S. 73 ff.).

LAU

Die Längsschnittstudie LAU (Aspekte der Lernausgangslage und der Lernentwicklung) untersuchte in einem zweijährigen Turnus die Lernstände und -entwicklungen von einem Großteil der Hamburger Schüler von der fünften bis zur dreizehnten Klasse. Die teilnehmenden Schulen erhielten klassenbezogene Rückmeldungen, verbunden mit einer sozialen Bezugsnorm in Form eines fairen Vergleichs des Gesamtwerts. Hieran schloss sich eine qualitative Rezeptionsstudie an, in der acht gymnasiale Schulleitungen zu der Nutzung der Rückmeldungen interviewt wurden. Die Rückmeldung wurde als nützlich empfunden und fungierte als Anlass für Ursachenreflexionen. Dies regte zugleich Diskussionen über didaktische und methodische Fragen des Unterrichts zwischen den beteiligten Lehrkräften an. Die Reflexion wurde umso intensiver durchgeführt, je stärker die Ergebnisse zwischen Parallelklassen oder in einzelnen Inhaltsbereichen voneinander abwichen. Eine systematische Auseinandersetzung mit den Ergebnissen auf Gremien- oder Konferenzebene fand jedoch nicht statt. Konsequenzen für die weitere Unterrichtsentwicklung wurden weniger abgeleitet (vgl. Klug & Reh, 2000, S. 17 ff.).

QuaSUM

Die Untersuchung QuaSUM (Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik) in Klasse 5 und 9 in Brandenburg beschäftigte sich mit inner- und außerschulischen Lernbedingungen und sollte dem Ministerium Steuerungsdaten für Unterstützungssysteme und die Lehrplanentwicklung liefern. Damit die Schulen die Ergebnisse aus den Tests zugleich auch für die Qualitätsentwicklung nutzen konnten, wurden schulbezogene Berichte konzipiert, in denen die Klassenergebnisse nach verschiedenen Kriterien aufgeschlüsselt und in einen Vergleich mit den Gesamtdurchschnittswerten der gleichen Schulform gesetzt wurden (vgl. Peek, 2004a, S. 21 ff.).

Im Rahmen der Rezeptionsstudie QuaSUM 2 erhielten alle an dem Test teilnehmenden Schulen (Fachkonferenzleiter, Klassenleiter, Mathematikfachlehrer) einen teilstandardisierten Fragebogen. Der Rücklauf betrug 74,4 Prozent der 163 angeschriebenen Schulen. Knapp die Hälfte der befragten Fachlehrer hielten die Tests für die unterrichtliche Arbeit für wenig nützlich und waren der Ansicht, dass sie nur Unruhe in die Schulen hereinbrächten (vgl. Peek, 2004b, S. 91). Die Funktion der Tests sahen sie vor allem als ein externes Instrument zur Selbstreflexion. Zudem konnte festgestellt werden, dass die Lehrpersonen den Tests insgesamt sowie einer Veröffentlichung der Daten wesentlich skeptischer gegenüber standen als die Schulleitungen (vgl. ebd., S. 93). Dennoch gaben über 90 Prozent der Fachlehrer an, sich mit den Ergebnissen auseinandergesetzt zu haben. Von besonderem Interesse waren hierbei Standards, Vergleiche mit eigenen Einschätzungen sowie die Leistungsheterogenität in der Lerngruppe (vgl. ebd., S. 105 f.).

Der kommunikative Austausch erfolgte vorrangig fachbezogen mit anderen Mathematiklehrkräften und in Fachkonferenzen und bezog sich auf Handlungsstrategien bezüglich der Unterrichtsgestaltung und Standardsicherung (vgl. ebd., S. 104). Auf Gesamtkonferenzen wurden hingegen organisatorische Aspekte besprochen und die Ergebnisse lediglich bekannt gegeben (vgl. ebd., S. 106). Des Weiteren konnte ermittelt werden, dass eine positive Einschätzung der vorhandenen Kooperationsstrukturen und des Schulklimas die Intensität der Reflexionsphase befördern (vgl. ebd., S. 110).

Dennoch konnten nur wenige klar definierbare Handlungen, welche aus den Testergebnissen abgeleitet wurden, eruiert werden. Lediglich 17 Prozent der Fachlehrer der neunten Klassen gaben an, mindestens eine Konsequenz tatsächlich umgesetzt zu haben. Die Aktionen bezogen sich überwiegend auf die methodische und didaktische Unterrichtsplanung (vgl. ebd., S. 98 ff., 110). Somit ist resümierend festzuhalten, dass zwar 90 Prozent der Fachlehrer eine Reflexion bezüglich der Testresultate vornahmen, jedoch bei etwa 80 Prozent der Nutzungsprozess an dieser Stelle abbrach und keine spezifischen Handlungen erfolgten.

MARKUS

Die flächendeckende Studie MARKUS (Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext) in den achten Klassen ermittelte ebenfalls die Lernstände im Fach Mathematik. Als Rückmeldung erhielten die Lehrkräfte die Leistungsergebnisse, verbunden mit einem Kontextprofil der entsprechenden Klasse.

In der zugehörigen Rezeptionsstudie WALZER gaben 81 Lehrpersonen über einen Fragebogen Auskunft zu den Auswirkungen der Tests und den Einflussfaktoren auf die Nutzung (vgl. Schrader & Helmke, 2004, S. 141 ff.). Die Ergebnisse hinsichtlich der Rezeption waren eher ernüchternd. 55 Prozent der Lehrkräfte waren der Auffassung, dass ihnen externe Tests für ihre unterrichtliche Arbeit wenig nützen würden (vgl. ebd., S. 141 ff.). Knapp die Hälfte der Befragten hatte erwogen, Maßnahmen zu ergreifen, jedoch nur 20 Prozent setzten hingegen die Handlungen auch in der Praxis um. Bei einem Großteil der Lehrpersonen blieben die Tests somit ohne direkte Wirkung auf die Unterrichtsentwicklung. Dennoch kristallisierte sich folgender Zusammenhang heraus: Je größer die Unterstützung durch den schulischen Kontext wahrgenommen wird und je positiver die Innovationsfreudigkeit der Lehrpersonen sowie der Schulleitung externen Leistungsmessungen gegenüber ist, desto größer ist die Wahrscheinlichkeit, dass Maßnahmen zur Qualitätsentwicklung ergriffen werden (vgl. ebd., S. 143, 156). Als repräsentativ und tragfähig können diese Ergebnisse jedoch nicht eingestuft werden, da der Rücklauf von 1.876 angeschrieben Personen lediglich 4,7 Prozent betrug (vgl. ebd., S. 145).

5.7.2 Rezeptionsergebnisse zu individualdiagnostischen Tests

Klassencockpit

Das schweizerische Testkonzept Klassencockpit ist ein Modulsystem, welches auf Basis der Freiwilligkeit in den Klassenstufen 3 bis 9 dreimal jährlich eingesetzt wurde. Mithilfe von Klassencockpit erhielten die Lehrpersonen individualdiagnostische Daten für ihre Lerngruppe und konnten anhand der Ergebnisse eine Standortbestimmung vornehmen sowie die Lernentwicklung optimieren. In der zugehörigen quantitativen Rezeptionsstudie mit einer Rücklaufquote von knapp 100 Prozent konnte ermittelt werden, dass bei 96 Prozent der Lehrpersonen tatsächlich eine Verortung der Schülerleistungen im Sinne einer Standortbestimmung vorgenommen wurde. Dabei äußerte ein Großteil, die Rückmeldung als informativ, verständlich und nützlich zu empfinden. Nur knapp die Hälfte der Befragten gab jedoch an, den Unterricht auf Grundlage der Ergebnisse weiterentwickelt zu haben, so dass bei etwa 50 Prozent der Befragten keine Maßnahmen abgeleitet wurden. 28 Prozent der Lehr-

kräfte verwendeten die Daten zudem, um ihre eigenen Einschätzungen bezüglich der Schülerleistungen anzupassen (vgl. Moser & Keller, 2002).

Check 5

Als ein weiteres individualdiagnostisches Testinstrument ist Check 5 anzuführen, an dem 140 fünfte Klassen im schweizerischen Kanton Aargau teilnahmen (vgl. Tresch, 2007, S. 158). Die involvierten Lehrpersonen erhielten anschließend als Rückmeldung sowohl Klassen- als auch individuelle Schülerergebnisse (vgl. ebd., S. 242). Tresch begleitete das Testinstrument mit einer verpflichtenden Rezeptionsstudie, welche aus drei schriftlichen Befragungen bestand. Die erste erfolgte gleichzeitig mit der Ausschreibung von Check 5 und zielte auf die Einstellungen der Lehrpersonen ab. Die zweite Befragung wurde mit dem Erhalt der Ergebnisrückmeldung durchgeführt und umfasste die Nutzungsphasen Rezeption und Reflexion sowie die Intention von Maßnahmen. Fünf Monate später folgte die dritte Befragung, bei welcher die Umsetzung der geplanten Handlungen thematisiert wurde (vgl. ebd., S. 158 ff.).

Bezüglich der Rezeption konnte festgestellt werden, dass die Rückmeldungen grundsätzlich verstanden wurden (vgl. ebd., S. 244). Die Lehrpersonen erwarteten generell eher durchschnittliche Leistungen ihrer Schüler, wobei es ihnen leichter fiel, individuelle Ergebnisse zu beurteilen, als die Klassenleistung im sozialen Vergleich (vgl. ebd., S. 252, 257). Theoretische Annahmen bestätigend (vgl. Abschnitt 5.5) attribuierten sie defizitäre Resultate vorwiegend external. Jedoch wurde die Annahme einer vorrangig internalen Ursachenzuschreibung bei überdurchschnittlichen Ergebnissen nicht belegt, da hierbei ebenfalls external attribuiert wurde (vgl. ebd., S. 259 ff.). Die Rückmeldungen wurden besonders intensiv mit Eltern, Schülern und Kollegen kommuniziert, wobei sich nur etwa 40 Prozent der Lehrpersonen mit der Schulleitung über die Ergebnisse austauschte. Tresch schließt daraus, dass es den Lehrpersonen umso wichtiger erscheine, mit Personengruppen transparent über die Resultate zu diskutieren, je direkter diese von den Ergebnissen betroffen sind (vgl. ebd., S. 272 ff.).

Bezüglich der Nutzungsphase „Aktion“ konnte eine immense Fülle an Maßnahmen festgehalten werden, wobei lediglich sechs Prozent der Befragten keine Konsequenzen aus den Rückmeldungen ableiteten. Daher wurde angenommen, dass die Aktionsphase als gelungen zu betrachten ist (vgl. ebd., S. 278 ff.). Es wurde jedoch keine Differenzierung zwischen dem Umfang, der Bedeutung und der Wirksamkeit der verschiedenen Maßnahmen vorgenommen. Dennoch ist es beachtlich, dass 86 Prozent der geplanten Handlungen tatsächlich teilweise oder vollständig umgesetzt wurden (vgl. ebd., S. 284). Darüber hinaus konnte

konnotiert werden, dass jüngere Lehrkräfte quantitativ tendenziell mehr Maßnahmen ergriffen als Kollegen mit längerer Berufserfahrung. Letztere Gruppe konnte hingegen bei der gelungenen Umsetzung dieser Handlungsstrategien eine positivere Quote aufweisen (vgl. ebd., S. 286).

Generell bewerteten die Befragten die Tests und die Rückmeldungen als eine wichtige bis sehr wichtige Informationsquelle für die Leistungsbeurteilung, die individuelle Förderung und die Unterrichtsreflexion (vgl. ebd., S. 300 f.). Die Beurteilung des Nutzens war jedoch im Verlauf der Rezeptionsstudie bis zur dritten Befragung leicht rückläufig (vgl. ebd., S. 311). Zudem ist die Hälfte der Lehrpersonen der Auffassung, dass aufgrund der Tests messbare Ziele ins Zentrum rücken und zu einseitig die fachlichen Leistungen geprüft würden. Knapp ein Drittel vertrat die Ansicht, dass die Tests zu einem Teaching to the Test führen würden, während etwa ebenso viele Check 5 als eine persönliche Kontrolle von außen wahrnahmen (vgl. ebd., S. 306).

BeLesen

Ein weiteres Testinstrument zur Individualdiagnostik ist die Längsschnittstudie BeLesen, bei der Berliner Erstklässler bis zum Ende der vierten Jahrgangsstufe halbjährlich getestet wurden. Die Lehrpersonen erhielten nach jedem Messzeitpunkt klassenbezogene sowie individuelle Informationen (vgl. Schneewind, Merkens, & Kuper, 2005, S. 86; Schneewind, 2006, S. 115). Schneewind begleitete BeLesen mit einer Rezeptionsuntersuchung, in der 56 Lehrkräfte insgesamt fünf Fragebögen vorgelegt wurden und zusätzlich zehn Personen problemzentriert interviewt wurden (vgl. Schneewind, 2006, S. 116).

Im Zentrum der Studie stand der Einfluss des Inhalts und der Gestaltung der Rückmeldung auf den Nutzungsprozess. Die Befragten beurteilten die Rückmeldungen größtenteils als verständlich und nützlich. Die Einschätzung der Bedeutung der Informationen für die eigene Arbeit nahm allerdings im Verlauf der Rezeptionsstudie ab (vgl. Schneewind, 2007b, S. 148). Zudem empfanden die Lehrpersonen die ermittelten Schülerresultate überwiegend als zutreffend. Zu etwa einem Viertel ihrer Schüler erhielten sie neue diagnostische Informationen (vgl. ebd., S. 145). Insbesondere defizitäre Ergebnisse wurden reflektiert, wobei die Analyse umso intensiver erfolgte, je spezifischer die Information auf den einzelnen Schüler bezogen war (vgl. ebd., S. 7).

Die Ergebnisse wurden hauptsächlich innerhalb des Kollegiums besprochen; etwa 50 Prozent der Befragten kommunizierte auch mit den Schülern darüber und bei 40 Prozent waren die Rückmeldungen Diskussionsgegenstand auf einer Fachkonferenz (vgl. ebd., S. 141). Die abgeleiteten Konsequenzen betrafen meist die individuelle Förderung oder die Unter-

richtsplanung. Zur Leistungsbewertung verwendete der Großteil der Lehrpersonen die Resultate selten oder nie (vgl. ebd., S. 157 f.). Bei Analyse der getroffenen Maßnahmen stellte Schneewind fest, dass die Handlungen in der Regel zuvor schon im professionellen Selbstverständnis und im unterrichtlichen Agieren verankert gewesen und durch die Rückmeldungen lediglich verstärkt worden seien (vgl. ebd., S. 156, 161).Aus diesem Grund hätten die Lehrpersonen die Rückmeldungen nur kaum als einen innovativen Handlungsimpuls begriffen. Als hauptsächliche Gründe für einen vorzeitigen Abbruch des Nutzungsprozesses wurden die fehlende Zeit und die mangelnde Bereitschaft angegeben (vgl. ebd., S. 7).

5.7.3 Rezeptionsuntersuchungen zu den Vergleichsarbeiten

VERA

Als erste Rezeptionsuntersuchung zu den Vergleichsarbeiten soll an dieser Stelle die Studie von Koch, et al. vorgestellt werden, welche die Nutzung von der Vergleichsarbeiten im Grundschulbereich (VERA in Klassenstufe 4) erforschte, um darauf aufbauend eine Evaluation des Testinstruments vornehmen zu können. Hierbei wurden alle teilnehmenden Lehrpersonen gebeten, einen Online-Fragebogen zu beantworten. Die Rücklaufquote betrug 19,6 Prozent (vgl. Koch, Groß Ophoff, Hosenfeld, & Helmke, 2006, S. 191 f.).

Die Verständlichkeit und Nützlichkeit der Ergebnisse wurden insgesamt als positiv bewertet. Im Zentrum der Auseinandersetzung stand vor allem der kriteriale Vergleich anhand von Fähigkeitsniveaus und Kompetenzbereichen; ein sozialer Vergleich mit Parallelklassen wurde selten vorgenommen (vgl. ebd., S. 193). Überraschend ist die Angabe von nahezu der Gesamtheit aller Befragten, dass die Ergebnisse von ihren eigenen Einschätzungen abweichen würden. Bei der Analyse dieser Differenzen wurden vor allem externale Attribuierungen (Schüler, Testkonstruktion, sonstige externe Bedingungen) vorgenommen. Lediglich 11 Prozent der Lehrkräfte gaben als Ursache für die Resultate den eigenen Unterricht an. Dennoch führten die Rückmeldungen bei über der Hälfte der Befragten zu einer Selbstreflexion bezüglich der eingesetzten Unterrichtsmethoden (vgl. ebd., S. 195). Auffällig ist zudem die hohe kooperative Auseinandersetzung mit den Testergebnissen. Lediglich 2,5 Prozent der Lehrpersonen kommunizierten mit keinen weiteren Personengruppen, wie Kollegium, Fachkonferenzen, Eltern oder Schülern (vgl. ebd., S. 193 f.).

Auch in Bezug auf die Nutzungsphase der Aktion konnten positive Ergebnisse festgestellt werden: Knapp 94 Prozent der Befragten leiteten aus den Testresultaten Konsequenzen ab (vgl. ebd., S. 194). Jedoch umfasste ein Großteil dieser Maßnahmen die Wiederholung oder Vertiefung von Inhalten und Aufgabenformen, so dass bereits vorhandene Maßnahmen

verstärkt wurden. Innovative Maßnahmen, wie die Neuentwicklung von Materialien und Unterrichtskonzepten, ergriffen lediglich 40 Prozent der Lehrkräfte (vgl. Groß Ophoff, Hosenfeld, & Koch, 2007, S. 422). Generell betrafen die initiierten Handlungen meist die getestete Klasse, da es den Lehrern besonders schwer fiel, die gewonnen Erkenntnisse auch auf andere Lerngruppen zu übertragen (vgl. ebd., 2006, S. 195). Insgesamt konstatieren Koch, et al. (vgl. 2006, S. 193, 196) als Resümee, dass die einzelnen Nutzungsphasen Rezeption, Reflektion und Aktion gelungen seien. Als hindernde Bedingung für die Nutzung der Vergleichsarbeiten gaben die Lehrkräfte mangelnde zeitliche und organisatorische Ressourcen an und forderten für den Mehraufwand eine Form des Ausgleichs (vgl. Groß Ophoff, Koch, Hosenfeld, & Helmke, 2006, S. 35 f.).

Lernstandserhebungen in Nordrhein-Westfalen

Für den Sekundarbereich wurde für die Lernstandserhebungen in Nordrhein-Westfalen eine Rezeptionsstudie mittels eines Online-Fragebogens nach den Testdurchgängen 2005 und 2006 durchgeführt. Die Rücklaufquote betrug entsprechend der verschiedenen schulischen Personengruppen (Lehrkräfte, Fachkonferenzvorsitzende, Koordinatoren, Schulleitung) zwischen 25 und 52 Prozent (vgl. Bonsen, Büchter, & Peek, 2006, S. 136 ff.).

Generell konnte eine hohe Bereitschaft festgestellt werden, sich mit den Rückmeldungen auseinanderzusetzen, da nahezu die Gesamtheit aller Befragten die Resultate der Schüler rezipierte. Obwohl 2006 die Verständnisschwierigkeiten bei der Interpretation der Rückmeldung abnahmen, gaben dennoch 40 Prozent der Lehrpersonen an, dass ihnen die statistischen Angaben Probleme bereiten würden (vgl. Kühle & Peek, 2007, S. 432 ff.). Den Mathematiklehrkräften und Vorsitzenden der Mathematikfachkonferenz fiel die Rezeption insgesamt leichter, was mit der Aufgabentradition des Faches und dem geringerem Zeitaufwand für die Auswertung begründet wurde (vgl. Bonsen, Büchter, & Peek, 2006, S. 141). Die Reflexion der Daten erfolgte besonders intensiv, wenn die Resultate sich mit bereits vorhandenden diagnostischen Informationen verbinden ließen.

Positive Erkenntnisse wurden vor allem im Bereich der kooperativen Auseinandersetzung festgestellt: 87 Prozent der Befragten besprachen die Ergebnisse mit den Lehrkräften der Parallelklassen. Zudem fand in zwei Dritteln der teilnehmenden Schulen eine zentrale Koordination der Durchführung und Auswertung der Tests und der zugehörigen Rückmeldungen statt, währenddessen in 23 Prozent der Schulen sogar eigene Arbeitsgruppen für die Thematik Lernstandserhebung gebildet wurden (vgl. Kühle & Peek, 2007, S. 437). Jedoch kommunizierte nur etwa jede vierte Lehrkraft die Testresultate mit den Eltern (vgl. ebd., S. 432).

Auch die Aktionsphase trat bei einem Großteil der Befragten ein. Generell äußerten 75 Prozent der Lehrkräfte die Bereitschaft, Konsequenzen für den Unterricht zu ziehen. Diese Bereitschaft konnte größtenteils auch praktisch umgesetzt werden, indem 64 Prozent Veränderungen in ihrem Unterricht vornahmen. In dieser Kategorie wurden bei Schulleitungsmitgliedern noch höhere Werte gemessen (vgl. Bonsen, Büchter, & Peek, 2006, S. 143 f.). Jedoch umfassten die Maßnahmen oftmals eine verstärkte Konzentration auf die Testinhalte und -formate, währenddessen nur 13 Prozent veränderte Unterrichtsgestaltung und -methoden konnotierten (vgl. Kühle & Peek, 2007, S. 442).

Insgesamt empfanden 64 Prozent der Befragten die Tests und ihre Rückmeldungen für die eigene Arbeit als nützlich. Es ist hierbei anzumerken, dass entgegen den Ergebnissen anderer Rezeptionsstudien, in denen eine Abnahme in der Bewertung der Nützlichkeit über die Erhebungszeiträume hinweg zu konstatieren war, die Einschätzung der Nützlichkeit 2006 im Vergleich zum Vorjahr nicht unwesentlich gestiegen war (vgl. ebd., S. 434). Dennoch hatten 60 Prozent der Teilnehmer im Jahr 2005 Zweifel an der Angemessenheit des Verhältnisses des Aufwand zum Nutzen geäußert (vgl. Bonsen, Büchter, & Peek, 2006, S. 141 f.). Es waren in diesem Zusammenhang positivere Einstellungen bei Schulleitungen und Koordinatoren zu beobachten. Demnach steigt die Wahrnehmung der Nützlichkeit, je weiter die befragte Person vom direkten Unterrichtsgeschehen entfernt ist (vgl. ebd., S. 143). Die Rezeptionsstudie ergab des Weiteren, dass den größten Einfluss auf den Nutzungsprozess der wahrgenommene Grad der Bedeutsamkeit der Rückmeldungen hat, währenddessen die Intensität der Auseinandersetzung und die Bewertung der Verständlichkeit der Rückmeldung die Einschätzung der Wirksamkeit nicht direkt beeinflussen (vgl. ebd., S. 147).

Kompetenztests in Thüringen

Die Kompetenztests in Thüringen werden ebenfalls von einer Evaluationsstudie begleitet, welche nach jedem Testdurchlauf sowohl die Nutzung als auch die Akzeptanz des Testinstruments ermittelt. Hierfür können die schulischen Akteure einen Online-Fragebogen ausfüllen. Ergänzt wird die Untersuchung durch qualitative Schulleiter-, Lehrer- und Elternbefragungen (vgl. Nachtigall & Jantowski, 2007, S. 404). Zudem sollte an dieser Stelle bemerkt werden, dass das Thüringer Rückmeldesystem aus fünf Komponenten besteht: eine Sofortrückmeldung der Klassenergebnisse, einen Ergebnisbericht, welcher zudem den fairen Vergleich enthält, einen Schulbericht, einen Ergänzungsbericht sowie einen Landesbericht (vgl. ebd., S. 403).

Allgemein wurde stets eine große Akzeptanz der Tests und Rückmeldungen festgestellt (vgl. ebd., S. 404). Die Kompetenztests wurden vorrangig als Diagnose- und Vergleichsinstru-

ment wahrgenommen, aus denen sich Impulse für Förderung und Differenzierung im Unterricht ergeben (vgl. ebd., S. 405 ff.). Eine intensive Rezeption kam jedoch nur den Rückmeldekomponenten zuteil, welche klassenspezifische Informationen enthalten. Die Bewertung der Bedeutsamkeit der Rückmeldungen sinkt daher, je schulbezogener die Daten aufbereitet sind (vgl. ebd., S. 405, 408). Die Rückmeldungen regten zudem die innerschulische Diskussion an. 2007 besprachen 96 Prozent der Schulleiter die Testresultate mit dem Kollegium; größtenteils wurden sie auch in Gesamtkonferenzen thematisiert. Kommunikation mit außerschulischen Personengruppen wurde indes lediglich geringfügig vorgenommen. Auch die Annahme von Unterstützungsangeboten in Form einer Beratung fand nur selten statt (vgl. ebd., S. 406 f.; Nachtigall, 2010, S. 121).

Die ergriffenen Maßnahmen bezogen sich meist auf die individuelle Förderung und die Anpassung von Unterrichtsinhalten und Unterrichtsgestaltung (vgl. Nachtigall, 2009, S. 149). Lediglich 4 Prozent der befragten Lehrpersonen konnten 2007 keinen Nutzen in den Kompetenztests erkennen. Beim Längsschnittvergleich der Ergebnisse konnte festgestellt werden, dass zu Beginn die Akzeptanz und der innovative Umgang mit den gewonnenen Informationen mit wachsender Erfahrung mit dem Testinstrument anstieg (vgl. Nachtigall & Jantowski, 2007, S. 408). Erstmals 2008 sank die Einschätzung der Nützlichkeit deutlich im Vergleich zum Vorjahr, was jedoch mit der Testung neuer Kompetenzbereiche begründet wurde (vgl. Nachtigall, 2009, S. 136). Zudem gilt der wahrgenommene Mehraufwand als ein zentraler Faktor für die Akzeptanz der Leistungsmessung, währenddessen die Potenziale für die Unterrichtsentwicklung und Förderung noch nicht zufriedenstellend genutzt werden (vgl. ebd., S. 151).

Wie auch bei anderen Rezeptionsstudien konnten 2009 und 2010 bei den Schulleitungsmitgliedern positivere Einstellungen festgestellt werden (vgl. ebd., S. 146, 151; Nachtigall, 2010, S. 123). Des Weiteren waren Unterschiede in der Nutzung zwischen den verschiedenen Schulformen erkennbar: Während in Gymnasien und Regelschulen die Kompetenztests vorrangig zur Leistungsdiagnostik verwendet wurden, wurde in den Grundschulen ein verstärkter Nutzen für die eigene Unterrichtsgestaltung, die innerschulische pädagogische Diskussion und die Qualitätsentwicklung wahrgenommen (vgl. Nachtigall, 2009, S. 148). In Grundschulen wurden zudem eher innovative Maßnahmen aus den Rückmeldungen abgeleitet als in anderen Schulformen. Auch eine Kommunikation der Ergebnisse mit den Eltern erfolgte verstärkt in Grundschulen (vgl. ebd., S. 149).

Vergleichsarbeiten in Baden-Württemberg

Eine weitere Rezeptionsstudie untersuchte die Nutzung der Vergleichsarbeiten in den Klassenstufen 2 und 6 in Baden Württemberg. Hierbei ist anzumerken, dass die Auswertung und Interpretation der Testresultate den Lehrpersonen gänzlich selbst überlassen blieb. Auch eine soziale Bezugsnorm mittels Durchschnittslandeswerten in Form eines fairen Vergleichs wurde nicht zur Verfügung gestellt. Des Weiteren ist der Unterstützungsgrad bei der Auswertung und Nutzung der Ergebnisse durch administrative Stellen in Baden-Württemberg als vergleichsweise niedrig einzuschätzen (vgl. Maier, 2008b; 2008c, S. 67).

Die Rezeptionsuntersuchung von Maier enthielt mehrere Komponenten. 2006 wurden zum einen 570 Lehrpersonen unterschiedlicher Schulformen angeschrieben und um Teilnahme an einem Fragebogen gebeten. Die Rücklaufquote betrug 54 Prozent (vgl. Maier, 2007, S. 5). Die Mehrheit der Befragten bezweifelte hierbei die Nützlichkeit der Vergleichsarbeiten, da den Testergebnissen nur sehr eingeschränkt differenzierte Lerndiagnosen zu entnehmen seien. Je stärker der förderdiagnostische Nutzen wahrgenommen wurde, desto größer war die allgemeine Akzeptanz des Testinstruments. Als signifikanter Haupteffekt konnte die Schulform ermittelt werden, wobei positive Einschätzungen in Hinblick auf die Bedeutsamkeit der Tests vor allem durch Hauptschullehrer erfolgten. Des Weiteren wurde deutlich, dass Lehrpersonen mit großen Klassenstärken skeptischer dazu eingestellt waren, was mit dem verstärkten Mehraufwand bei der Korrektur begründet wurde. Zudem hatte ein hoher Anteil an Migranten in der Lerngruppe einen positiven Einfluss auf die Beurteilung des förderdiagnostischen Nutzens. Interessanterweise konnte jedoch kein Zusammenhang zwischen Akzeptanz und der Selbstwirksamkeitserwartung der Lehrkraft nachgewiesen werden. Maier (vgl. 2008a, S. 106 ff.) vermutet, dass letztere wohl erst in konkreten Handlungen sichtbar werde.

Zum anderen wurde der Fragebogen durch eine qualitative Befragung von insgesamt 56 Lehrkräften verschiedener Schulformen ergänzt (vgl. Maier, 2009, S. 133). Die kritische Haltung gegenüber den Vergleichsarbeiten konnte dabei bestätigt werden. Es erfolgte überwiegend keine Selbstreflexion. Auch die Kooperation verlief lediglich oberflächlich. 12 Lehrpersonen gaben an, bezüglich der Durchführung und Korrektur mit Kollegen Absprachen getroffen zu haben, während 19 Lehrkräfte von einem informellen Austausch der Ergebnisse berichteten. Eine systematische Kommunikation innerhalb der Schule wurde nicht vorgenommen. Die Ableitung von Handlungskonsequenzen war ebenfalls keine Selbstverständlichkeit: 20 der befragten Lehrpersonen nahmen keine Veränderungen in ihrem Unterricht vor. Die am häufigsten genannten Maßnahmen bestanden in verstärktem Üben

und der Übernahme von Aufgabenstellungen in Klassenarbeiten (vgl. Maier, 2008c, S. 69 f.; Maier, 2009, S. 134 ff.).

Die Rezeptionsstudie dehnte Maier im darauffolgenden Jahr aus, indem 825 Lehrkräfte verschiedener Schulformen nochmals einen Fragebogen erhielten. Der Rücklauf betrug 37 Prozent Zudem beteiligten sich auch 310 thüringische Lehrer, die an den dortigen Kompetenztests teilgenommen hatten, an der Untersuchung. Die Rücklaufquote erreichte hierbei 48 Prozent. Der Grund für den Vergleich der Rezeption zwischen den beiden Bundesländern waren deren konträr zueinander konzipierten Rückmeldesysteme. Während die Lehrpersonen in Baden-Württemberg die Auswertung selbst vornehmen mussten, wurden für die Kollegen in Thüringen mehrere ausführliche Rückmeldeberichte mit aggregierten Vergleichsdaten erstellt (vgl. Maier, 2008b). Die Tests wurden in Thüringen grundsätzlich in allen Kategorien als positiver, nützlicher und weniger als eine Belastung bewertet als in Baden-Württemberg. Während die kooperative Auseinandersetzung in Baden-Württemberg überwiegend aus informellen Gesprächen mit Kollegen der Parallelklassen bestand, fand sie in Thüringen verstärkt systematisch in Konferenzen statt. Zudem konnten die Erkenntnisse bezüglich des Einflusses der Schulform bestätigt werden: Die Tests wurden generell von den Hauptschullehrern als nützlicher wahrgenommen, während Gymnasiallehrkräfte eher eine selektionsdiagnostische Bedeutung mit den Tests verknüpften. Des Weiteren wurden insbesondere die Mathematiktests in allen Kategorien besser beurteilt als die Tests im Fach Deutsch (vgl. ebd., S. 465 ff.).

Zusammenfassend betrachtet ergeben die Ergebnisse aus den vorgestellten Rezeptionsstudien kein eindeutiges Bild. Den Testinstrumenten und Rückmeldungen selbst wird oftmals eine hohe Qualität bescheinigt, doch die Einschätzung bezüglich deren Nutzen variiert enorm. Dennoch lässt sich feststellen, dass die Ergebnisse der individualdiagnostischen Messinstrumente im Vergleich zu anderen Testkonzepten verstärkt als bedeutsamer und nützlicher bewertet werden. Auch konnte in den Rezeptionsstudien zu den Vergleichsarbeiten eine große allgemeine Akzeptanz konnotiert werden. Ausnahme bildet hierbei Baden-Württemberg, wobei dies in der Tat damit begründet werden kann, dass die Nutzung dort in einem wesentlich geringeren Maß durch ein Rückmeldekonzept beeinflusst und gesteuert wird.

Dennoch sind die Erkenntnisse bezüglich der Aktionsphase ernüchternd. Bei den Bildungsmonitoring-Studien LAU, QuaSUM und MARKUS brach der Nutzungsprozess stets mit der Reflektion ab. Positivere Ergebnisse sind bei individualdiagnostischen Tests und den Vergleichsarbeiten zu konnotieren. Doch auch hier können große Unterschiede festgestellt

werden: Während bei VERA 4 oder Check 5 fast die Gesamtheit der Befragten von Unterrichtsveränderungen berichtete, sind es bei den Vergleichsarbeiten in Baden-Württemberg weniger als die Hälfte der Befragten. Tendenziell wurden bereits vorhandene Unterrichtsinhalte und -methoden verstärkt, währenddessen innovative Maßnahmen zu einer Unterrichtsentwicklung eher selten eingeleitet wurden. Auch das Ausmaß der kooperativen Auseinandersetzung mit anderen Personengruppen unterschied sich zwischen den einzelnen Testkonzepten zum Teil fundamental voneinander, wie zum Beispiel zwischen den Vergleichsarbeiten in Baden-Württemberg und den Kompetenztests in Thüringen.

Zudem sollten die methodischen Vorgehensweisen der Rezeptionsstudien kritisch betrachtet werden. Ein Großteil der vorgestellten Untersuchungen wurden von den Entwicklerteams der Tests oder der Rückmeldungen selbst vorgenommen und nicht von unabhängigen Wissenschaftlern (vgl. Maier, 2010a, S. 113). Beim Großteil der Rezeptionsstudien wurden quantitative Fragebögen eingesetzt; lediglich bei LAU, BeLesen und den Vergleichsarbeiten in Baden-Württemberg enthielten die Untersuchungen auch qualitative Elemente. Mit den quantitativen Forschungsmethoden waren zum einen teilweise äußerst geringe Rücklaufquoten verbunden (vgl. Levin, 2009, S. 53). Zum anderen stellt sich auch die Frage nach der Belastbarkeit der Ergebnisse. Kohler & Schrader (vgl. 2004, S. 12 f.) führen hierzu kritisch an, dass zum Beispiel aus einer positiven Beurteilung der Verständlichkeit der Rückmeldung nicht hervorgeht, ob die Lehrkräfte tatsächlich die Informationen verstanden haben. Auch bedeute die Angabe von initiierten Maßnahmen im Unterricht nicht, dass die Veränderungen tatsächlich im intendierten Sinne auf die Qualität des Unterrichts wirken.

Generell liefern Ergebnisse aus den Fragebogenuntersuchungen nur wenige Erkenntnisse bezüglich der tatsächlichen Nutzungsprozesse in den Schulen. Aus diesem Grund bestehen Forschungsdesiderate insbesondere qualitativer Studien und Längsschnittuntersuchungen, in denen die subjektiven Handlungsstrategien der schulischen Akteure sichtbar werden (vgl. Hartung-Beck, 2009, S. 43). Die Fragestellung nach den tatsächlichen Effekten auf die Qualitätssicherung und -entwicklung ist bislang ungeklärt (vgl. Maier, 2008a, S. 101; Schneewind, 2007a, S. 369). Auch wurde ungenügend untersucht, welche Bereiche der schulischen Arbeit überhaupt von den Rückmeldungen beeinflusst werden können, von welchen Bedingungen die Nutzung abhängt und in welcher Weise dieser komplexe Wirkungszusammenhang durch zusätzliche Unterstützungsangebote befördert werden kann (vgl. Kohler & Schrader, 2004, S. 9; Maier, 2008a, S. 95, 101).

TEIL B - METHODISCHE BETRACHTUNGEN

6 Untersuchungsdesign

6.1 Einbettung der Untersuchung in den methodischen Kontext

Die Untersuchung ist als qualitative Studie dem Bereich der Bildungs-, Schul- und Evaluationsforschung zuzuordnen. Da die Frage der Nutzung der Vergleichsarbeiten in den Schulen im Zentrum der Betrachtungen steht, werden im Rahmen der Bildungsforschung Bildungsprozesse im Kontext von Staat und Gesellschaft analysiert. Spezifizierend gehört dies zu der Schul- und Unterrichtsforschung als Teil der Bildungsforschung. Es ist das Ziel, Erkenntnisse über die Handlungen und die zugehörigen Kausalzusammenhänge im Kontext der Nutzung der Vergleichsarbeiten auf Seiten der schulischen Akteure zu erhalten (vgl. Ackermann & Rosenbusch, 2002, S. 31). Die Untersuchung der Forschungsfragen (vgl. Abschnitt 1.2) verlangt zudem die Betrachtung der Auswirkungen der Tests und ihrer Rückmeldungen auf das Verhalten der Lehrpersonen und Schulleitungen im Bereich der Schulentwicklung. Aus diesem Grund ist die Studie ebenso in der Evaluationsforschung zu verorten.

Die qualitative Untersuchung konzentriert sich darauf, die Handlungen bei der Verwendung der Vergleichsarbeiten deutend zu verstehen und somit eine Konstruktion der Realität an den Schulen zu erklären. Dies impliziert eine Analyse und Interpretation der Ursachen, der beeinflussenden Faktoren sowie der Wirkungen der schulischen Handlungsprozesse in diesem Kontext (vgl. Gläser & Laudel, 2010, S. 25 f.). Der qualitativen Forschung werden dabei die folgenden vier methodologischen Prinzipien zugrunde gelegt (vgl. ebd.,S. 30 ff.):

1. *Prinzip des theoriegeleiteten Vorgehens*

   Die Ergebnisse aus der Untersuchung sollten an vorhandenes theoretisches Wissen zum Forschungsgegenstand anknüpfen und sogleich einen Erkenntnisfortschritt generieren.

2. *Prinzip der Offenheit*

   Das Untersuchungsdesign und die sich anschließende Auswertung sollten für unerwartete Informationen oder sich ergebende Dynamiken offen sein. Auf diese Weise wird das Problem reduziert, die empirischen Erkenntnisse zu starr vorgefertigten Kategoriensystemen zuzuordnen. Würde diesem Prinzip gänzlich Rechnung getragen werden, würden die Beobachtungen nicht durch theoretische Vorbetrachtungen strukturiert werden, sondern die Kategorien ergäben sich aus der Untersuchung heraus. Aufgrund des „Prinzips des theoriegeleiteten Vorgehens“ wird oftmals eine Kombination aus de-

duktiver und induktiver Handhabung gewählt, so dass dennoch eine Offenheit bei der Durchführung und Auswertung gewährleistet wird.

3. *Prinzip des regelgeleiteten Vorgehens*

   Die Durchführungsschritte der Studie und deren zugrunde liegenden Regeln sollten dokumentiert und transparent beschrieben werden.

4. *Prinzip vom „Verstehen" als Basishandlung*

   Das Verstehen der Kausalmechanismen der beobachteten Handlungen steht im Zentrum der qualitativen Forschung, so dass die Ergebnisse auf der Interpretation der Sachverhalte basieren.

Das unter Berücksichtigung dieser methodologischen Prinzipien gewählte Untersuchungsdesign der vorgenommenen qualitativen Studie wird im Folgenden detailliert vorgestellt. Dieses wurde aus einer Wechselwirkung zwischen den Forschungsfragen und den theoretischen Vorüberlegungen heraus entwickelt. Nach Abschluss der Konzeption des methodischen Vorgehens erfolgte die Datenerhebung anhand einer Dokumentenanalyse (vgl. Abschnitt 6.2) und einer qualitativen Interviewstudie (vgl. Abschnitt 6.3). Anschließend wurden die Informationen zielgerichtet ausgewertet (vgl. Abschnitt 7) und interpretiert, so dass die Beantwortung der Fragestellungen dieser Arbeit ermöglicht wurde. Die Abbildung 16 stellt diesen Prozess grafisch dar.

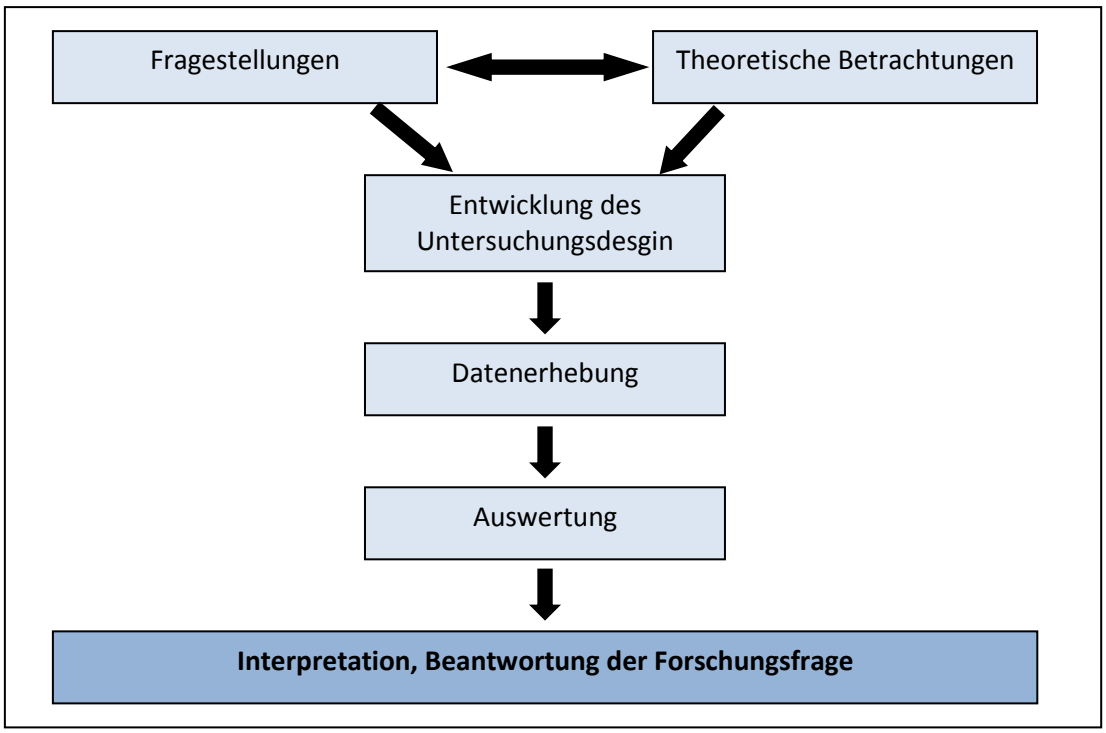

Abbildung 16: Struktur der qualitativen Forschung (vgl. Gläser & Laudel, 2010, S. 35)

6.2 Dokumentenanalyse

In Abschnitt 4.1 wurde ausführlich dargelegt, dass die Vergleichsarbeiten in jedem teilnehmenden Bundesland hinsichtlich ihrer Durchführungsmodalitäten, der Korrekturbestimmungen und ihres Rückmeldekonzepts voneinander variieren. Zudem wurde in einigen Bundesländern in der Sekundarstufe I lediglich eine Vergleichsarbeit in der Klassenstufe 8 implementiert, während in anderen Bundesländern (unter anderem auch in Hessen) zusätzlich in der sechsten Jahrgangsstufe die fachspezifischen Kompetenzen getestet werden.

Eine qualitative Untersuchung der Nutzung der Vergleichsarbeiten in den hessischen Gymnasien setzt eine Analyse der landesspezifischen Besonderheiten der Leistungsmessung in diesem Bundesland voraus. Aus diesem Grund hatte vor der Durchführung des zweiten Erhebungsinstruments, den qualitativen Interviews, eine detaillierte Erörterung der hessischen Vergleichsarbeiten zu erfolgen.

Dies wurde anhand einer Dokumentenanalyse vorgenommen, indem zum einen das hessische Rückmeldekonzept und dessen einzelne Bestandteile ein Untersuchungsaspekt waren. Wie in Abschnitt 5.5 beschrieben wurde, beginnt der Nutzungs- und Wirkungsprozess der Vergleichsarbeit primär mit dem Erhalt der Rückmeldung. Somit ist es für die Analyse der Nutzung der Schülerergebnisse durch die Lehrkräfte unerlässlich zu berücksichtigen, welche Informationen ihnen durch die Rückmeldung zur Verfügung gestellt und wie die Resultate der Tests aufbereitet werden.

Zum anderen wurde der Feedbackfragebogen analysiert, den die hessischen Lehrkräfte im Anschluss an den Erhalt der Rückmeldung online ausfüllen. Dieser soll dem für die Vergleichsarbeiten zuständigen Institut Informationen für die Weiterentwicklung der Leistungsmessung generieren. Da dieser Feedbackfragebogen eine institutionalisierte Möglichkeit darstellt, Hinweise von den Lehrpersonen bezüglich ihrer Nutzung der Vergleichsarbeiten zu erhalten, ist er für die Fragestellung dieser Untersuchung ebenfalls von Bedeutung.

Bei dieser Analyse der Rückmeldung und des Feedbackfragebogens wurden keine schulspezifischen Dokumente verwendet, welche anonymisierte Schülerdaten und -leistungen enthalten würden, sondern Musterrückmeldungen mit fiktiven Testergebnissen, welche durch das für Hessen zuständige Institut zur Verfügung gestellt wurden.

6.3 Teilstandardisierte Interviews als zentrale Erhebungsmethode

Aus der Präsentation der wesentlichen Ergebnisse aus bisherigen Forschungsstudien zur Nutzung der Vergleichungsarbeiten in Abschnitt 5.7 wurde die Dominanz quantitativer Untersuchungen gegenüber der qualitativen Forschung ersichtlich. Da sich die Fragestellungen dieser Arbeit ebenso wie einige andere Studien eng an dem Zyklenmodell von Helmke (2004) orientieren, würde eine weitere quantitative Untersuchung in diesem Kontext vermutlich keinen erkennbaren Mehrgewinn für den Forschungsstand generieren. Die vorliegende Arbeit dient vielmehr der Erweiterung zu bestehenden quantitativen Forschungsergebnissen, indem ein besonderer Schwerpunkt auf die Erklärung von Ursachen und Wirkungen von Nutzungshandlungen bei den schulischen Akteuren gelegt wird. Die zugehörigen beeinflussenden Faktoren sowie die Einstellungen, Deutungen und Sichtweisen der betreffenden Lehrpersonen und Schulleitungen können gezielter mithilfe einer qualitativen Untersuchung in Form von Interviews erfasst werden (vgl. Gläser & Laudel, 2010, S. 12).

Die Wahl des Interviews als zentrale Erhebungsmethode bietet mehrere Vorteile: Zum einen ermöglicht es eine höhere Fallorientierung, da tiefgründige Informationen zum Verhalten der Befragten und den damit verbundenen Haltungen und Emotionen generiert werden können. Dies impliziert auch wertvolles Wissen über die Hintergründe und die zugrundeliegenden Bedingungen von Verhaltensweisen. Durch differenziertere Äußerungen der Befragten, als es zum Beispiel mittels eines Fragebogens möglich wäre, kann das Wirkungsgefüge im Kontext der Nutzung der Vergleichsarbeiten ganzheitlicher und komplexer erfasst werden. Dies ermöglicht eine fundierte Bewertung und reduziert die Gefahr von Fehlschlüssen und Missinterpretationen. Durch den persönlichen Kontakt wird des Weiteren eine größere Offenheit in der Kommunikation erzeugt, was die Authentizität der Antworten der Befragten erhöht (vgl. Kuckartz, Dresing, Rädiker, & Stefer, 2008, S. 11 ff.).

Als Interviewform wurde das teilstandardisierte Interview ausgewählt, bei dem ein Leitfaden als Gesprächsgrundlage entwickelt wurde (vgl. Abschnitt 6.3.2). Bei dieser Interviewform werden die Handlungen des Interviewers durch den Leitfaden teilweise standardisiert, währenddessen es dem Befragten offen bleibt, wie er auf die Fragen antwortet. Die genaue Formulierung der Fragen und deren Reihenfolge im Interview können jedoch situativ variiert werden. Der Inhalt der Fragen des Leitfadens ergibt sich aus den Fragestellungen der Untersuchung sowie den theoretischen Vorbetrachtungen, so dass auf diese Weise ein theoriegeleitetes Vorgehen als methodologisches Prinzip qualitativer Methoden (vgl. Abschnitt 6.1) gewährleistet wird. Die flexible Handhabung des Leitfadens sowie die Gelegenheit, zu den Antworten des Gesprächspartners Nachfragen stellen zu können, ermöglichen eine Offenheit des gewählten Erhebungsinstruments sowie eine Konzentration auf das Ver-

stehen und Deuten der Interviewaussagen (vgl. Gläser & Laudel, 2010, S. 41 f.; Mayer, 2006, S. 36).

Der Leitfaden wurde mit einem Kurzfragebogen kombiniert, welcher relevante soziale Daten der Befragten erfasste und in Form eines Gesprächs die Dauer des Interviews zu stark ausgedehnt hätte. Der Kurzfragebogen beinhaltete folgende Kategorien:

- Geschlecht,
- Alter,
- Berufserfahrung (in Jahren)
- Unterrichtsfächer,
- Durchführung der Vergleichsarbeit im Unterrichtsfach X und der Klassenstufe Y,
- Anzahl der bereits durchgeführten Vergleichsarbeiten (in Jahren),
- Besuch von Fortbildungen zum Thema „Vergleichsarbeiten“, „Bildungsstandards“, „Kompetenzorientierung“,
- besondere Funktionen innerhalb der Schulorganisation.

Zudem wurden mit einigen Probanden kurze Nachgespräche telefonisch oder per E-Mail durchgeführt, wenn zum Zeitpunkt des Interviews der Nutzungsprozess nach Aussage des Befragten noch nicht abgeschlossen war.

6.3.1 Stichprobe

In Rahmen der Stichprobenauswahl wurde die Durchführung von Interviews an zwölf hessischen Gymnasien festgelegt. Die Einheitlichkeit der Schulform kann damit begründet werden, dass die Rahmenbedingungen an den Schulen zu verschieden und somit nicht mehr vergleichbar wären beziehungsweise zusätzliche Aspekte berücksichtigt werden müssten, die in dieser Untersuchung nicht erhoben werden konnten.

Die teilnehmenden Gymnasien wurden über ein Zufallsverfahren ausgewählt. Dabei wurde eine Liste aller hessischen staatlichen Gymnasien über ein computergestütztes Random-Verfahren in eine zufällige Rangfolge überführt. Anschließend wurden die Schulen der Reihenfolge nach kontaktiert. Dabei wurde zunächst ermittelt, ob erstens die Schule an den Vergleichsarbeiten teilnimmt und ob sie zweitens mit der Teilnahme an der Studie einverstanden ist. Auf diese Weise wurden die zwölf Gymnasium ausgewählt. Anschließend erhielten die Schulen Informationsmaterialien zu der Untersuchung sowie ein Einwilligungsformular. Die Auswahl der Gymnasien hätte zielgerichteter stattfinden können, wenn lediglich die an den Vergleichsarbeiten teilnehmenden Schulen randomisiert und kontaktiert

worden wären. Aufgrund von Datenschutzgründen war eine Auflistung der an den Tests teilnehmenden Schulen jedoch nicht zu erhalten. Selbstverständlich beruhte die Teilnahme an dieser Studie auf Freiwilligkeit.

Wie der Beschreibung der Stichprobenauswahl zu entnehmen ist, wurden bei der Auswahl der Schulen keine geographischen oder sonstige Faktoren berücksichtigt. Die Übersicht in Tabelle 4 zeigt die Verteilung der teilnehmenden Gymnasien nach den hessischen Schulamtsbezirken.

Schulamtsbezirk	Anzahl der Gymnasien
Landkreis Bergstraße und Odenwaldkreis	0
Landkreis Darmstadt-Dieburg und Stadt Darmstadt	1
Stadt Frankfurt am Main	1
Landkreis Fulda	2
Landkreis Gießen und Vogelsbergkreis	1
Landkreis Groß-Gerau und Main-Taunus-Kreis	0
Mainz-Kinzig-Kreis	0
Landkreis Hersfeld-Rotenburg und Werra-Meißner-Kreis	0
Hochtaunuskreis und Wetteraukreis	1
Landkreis und Stadt Kassel	1
Lahn-Dill-Kreis und Landkreis Limburg-Weilburg	1
Landkreis Marburg-Biedenkopf und Stadt Marburg	1
Landkreis Offenbach und Stadt Offenbach am Main	1
Rheingau-Taunus-Kreis und Stadt Wiesbaden	2
Schwalm-Eder-Kreis und Landkreis Waldeck-Frankenberg	0

Tabelle 4: Verteilung der teilnehmenden Gymnasien nach Schulamtsbezirken

Der Hauptadressat der Vergleichsarbeiten sind die teilnehmenden Lehrkräfte sowie die Schulleitungen. Dementsprechend wurden die Probanden für die Interviews aus diesen beiden Personengruppen ausgewählt. In jedem teilnehmenden Gymnasium wurde jeweils ein Interview mit einem Schulleitungsmitglied sowie mit einer Lehrkraft, deren Klasse getestet wurde, durchgeführt. Die Lehrpersonen sind als die zentralen Ansprechpartner anzusehen, da sie unmittelbar mit der Nutzung der Vergleichsarbeiten konfrontiert waren und ihre Erfahrungen und Einstellungen dem Testinstrument gegenüber berichten konnten. Die Schulleitungen wurden in der Untersuchung berücksichtigt, da die Vergleichsarbeiten nicht ausschließlich in der Unterrichtsentwicklung zur Qualitätssicherung und -entwicklung beitragen sollen, sondern gleichermaßen auch die Organisations- und Personalentwicklung anstoßen sollen. Die Schulleitung kann Informationen dazu liefern, inwiefern die Tests in einem solchen größeren Kontext von Schulentwicklung innerhalb der Schule genutzt wurden. Zudem hat die Schulleitung einen Überblick über die generelle Nutzungsintensität und

die Akzeptanz der Vergleichsarbeiten innerhalb des Kollegiums und kann dies entsprechend im Interview kommunizieren.

Da in jedem an der Untersuchung beteiligten Gymnasium nur zwei Interviews durchgeführt wurden, kann kein vollständiger Einblick in die Nutzung der Vergleichungsarbeiten an der jeweiligen Schule gewonnen werden. Hierzu wäre es notwendig, zumindest mit jeder am Test beteiligten Lehrperson ein Gespräch zu führen. Dies hätte jedoch eine zu hohe organisatorische und zeitliche Belastung der Schule zur Folge gehabt. Des Weiteren wäre der Vergleich der Aussagen zwischen den Schulen auf diese Weise nicht aussagekräftiger geworden, da aufgrund der Freiwilligkeit bei der Teilnahme an den Vergleichsarbeiten in Hessen (vgl. Abschnitt 8.2.1) die Beteiligung in jeder Schule unterschiedlich groß ist. Daher wurde eine einheitliche Struktur gewählt, so dass in jedem Gymnasium ein Vertreter der Schulleitung und eine beteiligte Lehrkraft befragt wurden. Somit wurden in den zwölf Schulen insgesamt 24 Personen befragt.

Dabei trat mehrfach der Fall ein, dass das Schulleitungsmitglied ebenfalls in der Rolle der durchführenden Lehrperson der Vergleichsarbeiten agierte. In diesem Fall konnten diese zusätzlichen Erfahrungen zusätzlich in der Auswertung der Interviews genutzt werden, so dass insgesamt 19 Personen befragt wurden, die an den Tests als Lehrkräfte selbst teilnahmen.

Die Vergleichsarbeiten werden in Deutsch, Mathematik und der ersten Fremdsprache durchgeführt. Eine Auswahl der beteiligten Lehrkräfte nach Fächern wurde nicht vorgenommen, da das Rückmeldekonzept für jedes Fach identisch ist. Es war daher davon auszugehen, dass standardisierte Faktoren wie der Aufbau der Rückmeldung keine Unterschiede im Nutzungsverhalten zwischen den verschiedenen Fächern zur Folge haben. Vielmehr war es für die Studie von Interesse, Aussagen zu den Vergleichsarbeiten in verschiedenen Fächern zu erhalten, um auf diese Weise die Konzeption der Tests im Fächervergleich als einen Einflussfaktor auf die Nutzungsintensität zu untersuchen.

Des Weiteren kann zwischen den Vergleichsarbeiten in der Klassenstufe 6 und 8 in Hessen differenziert werden, da VERA 6 nicht durch das IQB entwickelt wird (vgl. Abschnitt 4.1) Auch dieses Kriterium blieb unberücksichtigt bei der Auswahl der Probanden, da die Rückmeldungen von VERA 6 und VERA 8 identisch sind und die vorangegangen Argumente bezüglich der verschiedenen Fächern auch hier Anwendung finden.

Eine Übersicht zu den erfassten Personenmerkmalen der einzelnen Befragten kann der Tabelle 5 und der Tabelle 6 entnommen werden.

.

		Schule 1		Schule 2		Schule 3		Schule 4		Schule 5		Schule 6	
		Lehrkraft 1	Schulleitung 1	Lehrkraft 2	Schulleitung 2	Lehrkraft 3	Schulleitung 3	Lehrkraft 4	Schulleitung 4	Lehrkraft 5	Schulleitung 5	Lehrkraft 6	Schulleitung 6
Geschlecht	w/ m	w	m	m	m	w	w	m	w	w	w	w	m
Alter	in Jahren	58	58	61	54	51	40	35	58	38	46	28	58
Berufserfahrung	in Jahren	30	30	35	29	23	13	6	30	10	13	3	32
Besondere Funktionen			Stellv. Schulleiter	Ausbilder	Schulleiter		erweiterte Schulleitung		erweiterte Schulleitung, Koordinatorin		erweiterte Schulleitung		erweiterte Schulleitung
Teilnehmende Klasse	Kl. 6/ Kl. 8	Kl. 6	Kl. 6	Kl. 6		Kl. 6	Kl. 8	Kl. 6	Kl. 8	Kl. 6	Kl. 6	Kl. 8	
Teilnehmendes Fach	D/ M/ 1. FS	D und 1. FS	M	D		D	M	M und 1. FS	D	1. FS	D	D	
Anzahl der Teilnahmen	in Jahren	1	1	1		1	1	1	1	1	1	1	
Fortbildung zur Thematik	ja/ nein	ja	nein	ja	ja	ja	ja	ja	ja	nein	ja	nein	nein
Teilnahme der Schule im Fach Deutsch	Kl. 6/ Kl. 8	Kl. 6		Kl. 6 und Kl. 8		Kl. 6		Kl. 8		Kl. 6 und Kl. 8		Kl. 6 und Kl. 8	
Teilnahme der Schule im Fach Mathematik	Kl. 6/ Kl. 8	Kl. 6		Kl. 6 und Kl. 8		Kl. 6 und Kl. 8		Kl. 6		Kl. 6		Kl. 6	
Teilnahme der Schule in der 1. Fremdsprache	Kl. 6/ Kl. 8	Kl. 6		Kl. 6 und Kl. 8				Kl. 6		Kl. 6			
Anzahl der Teilnahmen der Schule	in Jahren	2		3		3		2		3		3	

Tabelle 5: Übersicht über die Probanden und Schulen - Teil 1

		Schule 7		Schule 8		Schule 9		Schule 10		Schule 11		Schule 12	
		Lehrkraft 7	Schulleitung 7	Lehrkraft 8	Schulleitung 8	Lehrkraft 9	Schulleitung 9	Lehrkraft 10	Schulleitung 10	Lehrkraft 11	Schulleitung 11	Lehrkraft 12	Schulleitung 12
Geschlecht	w/ m	w	m	w	w	m	m	m	m	m	w	w	m
Alter	in Jahren	32	51	61	54	55	46	42	50	57	46	28	62
Berufserfahrung	in Jahren	6	24	37	20	25	15	14	28	31	15	3	36
Besondere Funktionen		Fachsprecherin	erweiterte Schulleitung	Fachsprecherin	erweiterte Schulleitung, Koordinatorin	Fachsprecher	erweiterte Schulleitung	Fachsprecher	erweiterte Schulleitung		erweiterte Schulleitung, Koordinatorin		Schulleiter
Teilnehmende Klasse	Kl. 6/ Kl. 8	Kl. 6		Kl. 6	Kl. 6	Kl. 6	Kl. 6	Kl. 8		Kl. 8		Kl. 8	Kl. 8
Teilnehmendes Fach	D/ M/ 1. FS	1. FS		1. FS	M	M	D	1. FS		1. FS		M	M
Anzahl der Teilnahmen	in Jahren	2	0	1	1	1	1	1	0	1	0	1	1
Fortbildung zur Thematik	ja/ nein	ja	ja	ja	nein	ja	ja	nein	ja	ja	ja	ja	ja
Teilnahme der Schule im Fach Deutsch	Kl. 6/ Kl. 8	Kl. 6 und Kl. 8		Kl. 6		Kl. 8				Kl. 6			
Teilnahme der Schule im Fach Mathematik	Kl. 6/ Kl. 8	Kl. 6 und Kl. 8		Kl. 6		Kl. 6				Kl. 6		Kl. 6 und Kl. 8	
Teilnahme der Schule in der 1. Fremdsprache	Kl. 6/ Kl. 8	Kl. 6 und Kl. 8		Kl. 6		Kl. 6		Kl. 8		Kl. 8		Kl. 6 und Kl. 8	
Anzahl der Teilnahmen der Schule	in Jahren	3		2		3		2		2		3	

Tabelle 6: Übersicht über die Probanden und Schulen- Teil 2

Zuletzt ist auf zwei teilnehmende Gymnasien in besonderer Weise hinzuweisen: In beiden Fällen bestand die Schulleitung darauf, ein gemeinsames Interview mit der Lehrkraft durchzuführen. In den anderen Schulen wurden jeweils getrennte Gespräche vorgenommen. Das Interview bei diesen beiden Gymnasien fand hingegen mit beiden Gesprächspartnern gemeinsam statt. Dies stört die Einheitlichkeit des Forschungsdesigns. Zudem ist es unklar, ob beide Probanden aufgrund der Anwesenheit der jeweilig anderen Person ihre Einstellungen und Nutzungsprozesse tatsächlich ehrlich und umfassend erläutert haben. Andererseits ermöglicht ein solches Gespräch eine offene Diskussion mit mehr Teilnehmern zu dem Forschungsgegenstand, was wiederum neue Blickwinkel eröffnen kann. Zum anderen ist die Intensität er Kooperation bei der Nutzung der Vergleichsarbeiten bedeutsam. Dies kann mithilfe eines gemeinsamen Gesprächs eventuell präziser beobachtet und erfasst werden. Aus diesen Gründen heraus wurden die zwei Interviews dennoch in die Untersuchung aufgenommen.

6.3.2 Leitfaden des Interviews

Der Leitfaden diente als Grundgerüst für die Durchführung der Interviews. Die Konzeption der Fragen ergab sich aus den Fragestellungen der Arbeit sowie den theoretischen Vorüberlegungen. Angelegt wurde der Leitfaden für ein Gespräch der Dauer von dreißig bis vierzig Minuten. Da die Interviews mit den verschiedenen Probandengruppen (Lehrkräfte und Schulleitungsmitglieder) teilweise unterschiedliche Schwerpunkte implizierten, wurden zwei verschiedene Leitfäden entwickelt, welche lediglich in einigen Fragen übereinstimmen. Wie bereits erwähnt wurde, führten einige Schulleitungsvertreter die Vergleichsarbeit selbst in einer Klasse durch, so dass sie gewissermaßen eine Doppelrolle als Schulleitungsmitglied und als Lehrkraft in der Studie einnahmen. In diesen Fällen erfolgte eine Kombination der beiden Leitfäden.
Die Fragen der Leitfäden können der Abbildung 17 und
Abbildung 18 entnommen werden.

Leitfaden für die Lehrkräfte

Aus welchen Gründen nahmen Sie/ Ihre Schule an den Lernstandserhebungen teil?

Unterrichtsentwicklung

Wie beeinflusste die Durchführung einer Lernstandserhebung den Unterricht?

Welche inhaltlichen Bereiche der Ergebnisrückmeldung waren für Sie von besonderem Interesse?

Welche weiteren Informationen sollten Ihrer Meinung nach durch die Rückmeldung zur Verfügung gestellt werden?

Welche zusätzlichen Informationen oder klassenspezifischen Kontextfaktoren haben Sie bei der Auseinandersetzung mit den Ergebnissen herangezogen?

Inwiefern haben Sie die Ergebnisse zur Diagnostik der Schülerleistungen verwendet?

Inwiefern haben Sie Maßnahmen oder Konsequenzen aus den Ergebnissen für den Unterricht gezogen?

Wie gelang Ihnen die Umsetzung dieser Maßnahmen? Kann eine rückblickende Bewertung bereits vorgenommen werden?

Worin sehen Sie die Schwierigkeiten und Probleme bei der Nutzung der Lernstandserhebung?

Personalentwicklung

Inwiefern betrachten Sie die Ergebnisse als Feedback zu Ihrer eigenen Arbeit?

Wie messen Sie sonst – die Lernstandserhebung ausgenommen - Ihren Unterrichtserfolg?

Können Sie für sich selbst, bedingt durch die Nutzung der Lernstandserhebung, einen Kompetenzzuwachs in Ihrer Lehrerprofessionalität feststellen?

Organisationsentwicklung

In welcher Weise fand eine schulinterne Kommunikation über die Ergebnisse statt?

Wie schätzen Sie das Kooperations- und Innovationsklima an Ihrer Schule ein? Sind in diesem Zusammenhang Veränderungen durch die Lernstandserhebung festzustellen?

Mit welchen außerschulischen Personen/ Institutionen kommunizierten Sie die Ergebnisse?

In welchen Aspekten und Nutzungsphasen würden Sie Unterstützung von außen für sinnvoll erachten?

Fazit: Welches Potenzial bietet die Lernstandserhebung für die Arbeit in Schule und Unterricht?

Abbildung 17: Interviewleitfaden für Lehrkräfte

Leitfaden für Schulleitungsmitglieder

Aus welchen Gründen nahm Ihre Schule an den Lernstandserhebungen teil?

Unterrichtsentwicklung

Wie können die Tests und die Rückmeldungen für den Unterricht genutzt werden?

Inwiefern haben Sie Kenntnis zu Maßnahmen und Konsequenzen, die von Kollegen/innen aus den Ergebnissen abgeleitet und umgesetzt wurden?

Personalentwicklung

Inwiefern betrachten Sie die Rückmeldung als ein Feedback für die pädagogische Arbeit?

Inwiefern haben die Lernstandserhebungen eine grundsätzliche Auseinandersetzung über die schuleigenen Vorstellungen zum Lehren und Lernen angestoßen?

Wie schätzen Sie die Einstellungen Ihrer Lehrkräfte zu den Bildungsstandards und den Lernstandserhebungen ein?

Wie beurteilen Sie das Kooperations- und Innovationsklima an Ihrer Schule ein? Sind in diesem Zusammenhang Veränderungen durch die Lernstandserhebungen festzustellen?

Inwiefern beobachten Sie bei Ihren Kollegen/innen durch die Nutzung der Lernstandserhebung einen Lernzuwachs in deren Lehrerprofessionalität?

Organisationsentwicklung

Welche Personengruppen setzten sich mit der Auswertung der Lernstandserhebung auseinander?/ Welche Aspekte wurden wie intensiv diskutiert?/ Welche Rolle nahmen Sie als Schulleitung hierbei ein?

Inwieweit betrachten Sie die Testrückmeldung als eine Möglichkeit, um nachhaltig die qualitative Weiterentwicklung der eigenen Schule voranzutreiben?

Inwiefern können sie einen Zusammenhang zwischen den Lernstandserhebungen und weiteren Qualitätsmaßnahmen an Ihrer Schule sehen?

Welche Probleme oder Schwierigkeiten assoziieren Sie beim Umgang mit den Lernstandserhebungen?

Bei wiederholter Teilnahme: Inwiefern nehmen Sie Veränderungen im Umgang mit den Tests bei sich, den Lehrkräften und den Schülern/innen wahr?

Ist Ihnen ein Austausch der Erfahrungen und Ergebnisse aus den Lernstandserhebungen über die Kollegiums- und Schulgrenze hinaus bekannt? Wie stellte sich dieser dar?

Für wie hilfreich schätzen Sie die derzeitigen externen Unterstützungsmöglichkeiten ein (Schulaufsicht, Handreichungen)? Welche Unterstützung wäre notwendig?

Fazit: Welches Potenzial bieten die Lernstandserhebungen für die Arbeit in Schule und Unterricht?

Abbildung 18: Interviewleitfaden für Schulleitungsmitglieder

6.3.3 Erhebungszeitraum

Die Vergleichsarbeiten werden in jedem Schuljahr Ende Februar/ Anfang März in den hessischen Schulen durchgeführt. Anschließend haben die Lehrkräfte circa drei Wochen Zeit für die Korrektur. Nach der Online-Ergebniseingabe erhalten sie eine Sofort-Rückmeldung. Der Ergebnisbericht, welcher den Vergleich mit den hessischen Mittelwerten enthält, wird den Schulen Ende Mai zur Verfügung gestellt. Da es für die Nutzung der Vergleichsarbeiten elementar ist, alle Bestandteile der Rückmeldung erhalten zu haben, konnten die Interviews erst im Anschluss an den Erhalt des Ergebnisberichts durchgeführt werden. Weil das Schuljahr 2009/2010 jedoch bereits Anfang Juli endete, eröffnete sich nur ein relativ kleines Zeitfenster im Juni 2010 für die Durchführung der Gespräche. Dies setzte voraus, dass Handlungen als Reaktion auf die Auswertung der Ergebnisse von den Befragten unmittelbar eingeleitet worden waren beziehungsweise die Probanden sich während des Interviewzeitraums in der Nutzungsphase befanden. Da dies nicht immer zutraf, wurden in einigen Fällen Nachgespräche zu Beginn des folgenden Schuljahres 2010/2011 im September 2010 durchgeführt. Ein anderer Zeitraum für die Interviews war aufgrund des Zeitpunkts des Rückmeldeerhalts und dem Ende des Schuljahres nicht möglich.

Die

Tabelle 7 stellt den Erhebungszeitraum übersichtlich dar.

	Vergleichsarbeit	**Erhebung**
Februar 2010	Durchführung der Vergleichsarbeit	
März 2010	Durchführung der Vergleichsarbeit/ Korrektur	Auswahl der teilnehmenden Gymnasien
April 2010	Erhalt der Sofort-Rückmeldung	
Mai 2010	Erhalt des Ergebnisberichts	Auswahl der Gesprächspartner
Juni 2010		Durchführung der Interviews
Juli 2010		Transkription der Interviews
August 2010		
September 2010		gegebenenfalls Nachgespräche (telefonisch/ per E-Mail)

Tabelle 7: Übersicht über den Erhebungszeitraum

7 Auswertungsmethode

7.1 Transkription

Um eine genaue Auswertung zu ermöglichen, wurden die Interviews nach ihrer Durchführung vollständig transkribiert. Da es für die Transkription keine allgemein akzeptierten Vorgaben gibt (vgl. Gläser & Laudel, 2010, S. 193), wurde ein eigene Arbeitsgrundlage entworfen, die dem Auswertungsziel gerecht wurde. Es erfolgte eine Transkription in der Standardorthografie, die eine Übertragung in ein normales Schriftdeutsch umfasste. Sprachliche Färbungen und Dialekte fanden keine Übernahme in die Transkriptionen, so dass die Verschriftlichung der Gespräche leicht geglättet wurde. Pausen und Betonungen wurden nur dann festgehalten, wenn sie für die Bedeutung der Aussage wesentlich waren. Zudem wurden die Informationen, auf die Identität des Probanden oder der zugehörigen Schule rückschließen könnten, anonymisiert (vgl. Kuckartz, 2010, S. 38 ff.).

7.2 Qualitative Inhaltsanalyse

Als Auswertungsmethode der Interviews wurde die qualitative Inhaltsanalyse verwendet, welche sich an den Techniken nach Mayring (vgl. 2007, S. 24 ff.) orientiert. Diese Methode umfasst vier wesentliche Schritte:

1. Vorbereitung der Extraktion,
2. Extraktion,
3. Aufbereitung der Daten und
4. Auswertung der Daten.

Nach dieser Methode wird das Transkriptionsmaterial der Interviews anhand eines Analyserasters untersucht und wesentliche Informationen extrahiert. Diese werden anschließend unabhängig vom Originaltext weiterverarbeitet, so dass eine Informationsbasis entsteht. Die Quellenangabe wird mitgeführt, was bei Bedarf einen Rückbezug zum Ursprungstext ermöglicht. Das anzuwendende Analyseraster besteht aus Kategorien, welche zuvor aus den theoretischen Vorbetrachtungen und den Fragestellungen der Untersuchung heraus entwickelt werden. Dieses Kategoriensystem kann zum Beispiel hierarchisch oder gleichrangig aufgebaut sein. Bei der Extraktion ist es zwar möglich, Informationen mehreren Auswertungskategorien zuzuordnen, doch sollte das Kategoriensystem möglichst trennscharf konzipiert sein. Die Extraktion beinhaltet zudem eine Interpretation, indem die

transkribierten Interviews auf ihre Relevanz und ihren Informationsgehalt hin untersucht und beurteilt werden. Die qualitative Inhaltsanalyse stellt daher sowohl eine systematische als auch eine theoriegeleitete Auswertungsmethode dar.

Im Anschluss erfolgt die weitere Auswertung lediglich mit dem Extraktionsmaterial, welches auf Redundanzen und Widersprüche geprüft und somit weiter effizient reduziert wird. Dieser Schritt der Datenaufbereitung verbessert die Qualität der Auswertung und vereinfacht die Zusammenfassung und inhaltliche Strukturierung der Informationen (vgl. Gläser & Laudel, 2010, S. 229).

Im letzten Schritt der Auswertung können Kausalmechanismen auf verschiedenen Abstraktionsebenen identifiziert werden. Beispielsweise kann die individuelle Perspektive des Gesprächspartners wesentliche Aufschlüsse in Bezug auf die Forschungsfrage ergeben. Zum anderen können Ursachen für verschiedene Nutzungsprozesse und deren Wirkungen identifiziert und rekonstruiert werden. Des Weiteren ist eine Einbettung in den vorhandenen Forschungsstand vorzunehmen. Die Berücksichtigung dieser Ebenen in der abschließenden Auswertung ermöglicht letztlich die Beantwortung der Forschungsfrage (vgl. ebd. S. 246 ff.).

Da das Kategoriensystem nach Mayring vorab entwickelt wird und während der Auswertung nicht mehr veränderbar ist, kann das Problem entstehen, dass sich wesentliche Informationen den Kategorien nicht präzise zuordnen lassen (vgl. ebd. S. 198 f.). Aus diesem Grund wurde in dieser Interviewstudie in Anlehnung an Gläser und Laudel (vgl. ebd. S. 199 ff.) die Methode an das Design der Untersuchung angepasst, indem das Kategoriensystem während der Extraktion an die vorliegenden Informationen aus den Interviews modifiziert wurde. Die Auswertungsmethode erfüllt auf diese Weise das Kriterium der Offenheit, denn es ermöglicht die Zuordnung von unvorhergesehenen, aber relevanten Informationen zu geeigneten Kategorien. Es wurde daher eine Kombination aus induktiven und deduktiven Kategorien verwendet (vgl. Kuckartz, 2010, S. 62).

Dem Aufbau des Kategoriensystems liegt primär der Aufbau des Zyklenmodells nach Helmke (vgl. Abschnitt 5.5) zugrunde. Das Kategoriensystem ist infolgedessen teilweise hierarchisch aufgebaut, da beispielsweise die Nutzungsphasen aufeinander folgen. Die nachfolgende Tabelle 8 stellt das verwendete Kategoriensystem und sowie die Anzahl der zugeordneten Interviewaussagen dar.

Kategoriensystem	Anzahl der Zuordnungen
Rezeption	0
Testinhalte	6
Rückmeldung	63
Unterstützungsmaterialien	19
Reflexion	0
Reflexionsgegenstand	101
Diagnostische Kompetenz	45
Attribuierung der Ergebnisse	0
internale Attribuierung	24
externale Attribuierung	0
Attribuierung auf Schüler-/Klassenebene	17
Attribuierung auf Schulebene	3
Attribuierung auf Testebene	42
Kooperation und Kommunikation	0
Kollegium	0
inoffiziell	69
Fachkonferenz	46
Gesamtkonferenz	5
Schulleitung	39
Schüler	48
Eltern	34
außerschulisch	45
Aktion	0
Unterrichtsentwicklung	0
didaktisch	112
förderdiagnostisch	20
Bewertung und Benotung	21
Organisationsentwicklung	34
Personalentwicklung	53
generelle Schulentwicklung	25
Evaluation	0
Beeinflussende Bedingungen	0
individuelle Bedingungen	0
Motivation für die Teilnahme	0
intrinsische Motivation	43
extrinsische Motivation	26
Professionelle Selbst	37
Innovationsbereitschaft	17
Selbstreflexions- und Evalutationserfahrungen	22
schulische Bedingungen	0
Kooperationsbeziehungen	27
Innovationsklima	37
Entschädigung des Zeitaufwandes	17
externe Bedingungen	0
Lehrplan und sonstige Regularien	4
Test- und Rückmeldungszeitpunkt	33
Freiwilligkeit der Tests	16
Fortbildungssysteme	8
Bewertung des Lernstandserhebungen	0
Korrektur	83
Testgegenstand	67
Schwierigkeit	20
Rahmenbedingungen bei der Durchfürung	31
Rückmeldung	19
Benotung der Testergebnisse	22
Einschätzung des Nutzens der Lernstandserhebungen	0
Aufwand versus Nutzen	33
hessische Vergleichsarbeiten versus Lernstandserhebungen	0
Kooperation	33
Abschätzung des Nutzens	39

Tabelle 8: Kategoriensystem der qualitativen Inhaltsanalyse

Zuletzt sei darauf hingewiesen, dass bei der Auswertung der entnommenen Informationen keine Einzelfallanalyse oder Typisierung im Vordergrund stand, da dies für die Fragestellung nicht zielführend gewesen wäre und in Bezug auf die aktuelle Forschungslage nicht als notwendig zu betrachten ist (vgl. Abschnitt 5.7).

TEIL C - AUSWERTUNG DER UNTERSUCHUNG

8 Vergleichsarbeiten in Hessen: Analyse des Durchführungskonzepts der Lernstandserhebungen

Bevor die Ergebnisse der vorgenommenen Untersuchung zur Nutzung der Vergleichsarbeiten durch die schulischen Akteure am Beispiel des Bundeslandes Hessens vorgestellt werden, erfolgt eine detaillierte Betrachtung der landesspezifischen Besonderheiten bei der Durchführung der Tests sowie deren Einbindung in das Konzept der Standardisierung mittels der hessischen Bildungsstandards. Dies erscheint erforderlich, da die einzelnen Bundesländer für die Umsetzung und Implementierung der Bildungsstandards verantwortlich sind. Zudem wurde in Abschnitt 4.1 bereits erwähnt, dass signifikante Unterschiede in der Durchführung und Organisation der Vergleichsarbeiten sowie im Aufbau des jeweiligen Rückmeldekonzepts zwischen den einzelnen Bundesländern vorliegen.

8.1 Das Konzept der Standardisierung in Hessen

Als zuständiges hessisches Landesinstitut agiert das am 1. Januar 2013 eingerichtete Landesschulamt und Lehrkräfteakademie (kurz: LSA) in Wiesbaden. Zuvor war das Institut für Qualitätsentwicklung verantwortlich, welches als Evaluationseinrichtung zur Förderung der Qualitätsentwicklung im hessischen Bildungssystem am 1. Januar 2005 gegründet wurde und dessen Tätigkeitsfelder nun auf das LSA übertragen wurden.

Zu den Aufgaben des Instituts gehört unter anderem die Konzeption und fortwährende Weiterentwicklung des *Hessischen Referenzrahmens Schulqualität*. Dieser dient als „Grundlage für eine erfolgreiche Verständigung über die Güte von Schulen“ (Institut für Qualitätsentwicklung, 2008, S. 5) und verfolgt die Zielsetzung der Identifizierung von Stärken und Schwächen der Einzelschulen, so dass eine nachhaltige Qualitätsentwicklung angestoßen werden kann. Der Hessische Referenzrahmen Schulqualität wird als eine zentrale Bezugsgröße insbesondere für externe Evaluationen und Qualitätskontrollen verwendet, wie für die seit 2006 durchgeführten Schulinspektionen. Die Schulinspektion wird in einem Turnus von vier bis fünf Jahren vorgenommen, bei der mittels verschiedener Methoden, wie Dokumentenanalyse, Befragungen und Unterrichtshospitationen, die Einzelschule eine umfassende Analyse ihres Qualitätsstandes mit der Ausweisung ihrer Stärken und Defiziten sowie von Weiterentwicklungsmöglichkeiten erhält (vgl. Institut für Qualitätsentwicklung, 2011, S. 12 f.).

Für den Kontext der Vergleichsarbeiten in Hessen ist das explizite Aufgreifen der standardisierten Leistungsmessungen im Form externer Evaluationen im Hessischen Referenzrahmen Schulqualität von Bedeutung (vgl. Institut für Qualitätsentwicklung, 2008, S. 14). Demnach sollte sich in den Schulen eine Feedback-Kultur entwickeln, in der zur Verfügung gestellte Daten zur Schulqualität, wie die der Vergleichsarbeiten, für weiterführende interne Evaluationsprozesse verwendet sowie Handlungsbedarf für die Unterrichts- und Schulqualität abgeleitet werden (vgl. ebd., S. 32 f.). Dieses Qualitätskriterium wird bei den Schulinspektionen ebenfalls untersucht. Nach einer Auswertung der Inspektionen des Schuljahres 2009/2010 weisen die Schulen unabhängig von ihrer Schulform diesbezüglich noch relative Schwächen auf (vgl. Institut für Qualitätsentwicklung, 2011, S. 26 ff.). Besonders in Gymnasien wird der Handlungsbedarf aus externen Informationen nur teilweise formuliert und die Qualitätsentwicklung nur eingeschränkt systematisch und datenbasiert vorangetrieben (vgl. ebd., S. 54). Diese Ergebnisse lassen vermuten, dass die Vergleichsarbeiten in Hessen bislang noch keinen besonders großen Stellenwert für die Schulentwicklung in der schulischen Praxis einnehmen.

Neben den beiden Aufgabenbereichen der Weiterentwicklung des Hessischen Referenzrahmens für Schulqualität sowie der Schulinspektion stellt das LSA des Weiteren Instrumente zur Selbstevaluation für die Schulen bereit, die sich ebenfalls an den Qualitätskriterien des Referenzrahmens orientieren. Das Institut agiert zudem im Bereich der Qualitätssicherung in der Lehrerfortbildung, betreut schulische Modellprojekte und nimmt Wirkungsanalysen vor, welche Informationen zur Steuerung des Bildungssystems für das Hessische Kultusministerium generieren.

Außerdem ist das Institut für die Implementierung der Bildungsstandards und Kerncurricula in Hessen verantwortlich. Das vorangegangene Institut für Qualitätsentwicklung erhielt diesbezüglich vom Hessischen Kultusministerium den Auftrag, weiterführend zu den von der KMK entwickelten nationalen Bildungsstandards für die Hauptfächer und die Naturwissenschaften landeseigene Bildungsstandards für alle Fächer zu konzipieren. Diese beziehen sich zunächst auf die Primarstufe und die Sekundarstufe I. Langfristig sollen jedoch auch Bildungsstandards für die gymnasiale Oberstufe entwickelt werden. Die Erarbeitung der Standards erfolgte in interdisziplinären Teams aus Lehrpersonen aller Schulformen, Ausbildern der Studienseminare, Mitarbeitern des Instituts sowie wissenschaftlichen Gutachtern aus den Bereichen der Fachdidaktik und der Allgemeinen Didaktik (vgl. Höfer, Steffens, Diehl, Loleit, & Maier, 2010, S. 7).

Während in den nationalen Bildungsstandards der KMK konkrete Inhaltsbezüge hergestellt worden sind, orientieren sich die hessischen Bildungsstandards am Kompetenzbegriff und

verzichten auf konkrete inhaltliche Verbindlichkeiten. Stattdessen werden in Form von Inhaltsfeldern die inhaltlichen Zusammenhänge sichtbar, die jedoch nicht mit einzelnen Kompetenzen spezifisch verknüpft sind. Dies ermöglicht eine größere curriculare Freiheit für die praktische Umsetzung der Bildungsstandards (vgl. ebd., S. 8 f.). Des Weiteren sollen die Konzeptionen den Anforderungen an Bildungsstandards, wie sie in Abschnitt 2.2.2 beschrieben wurden, genügen, indem Einheitlichkeit, Prägnanz und Fokussierung als Maßgaben gestellt worden sind. Insbesondere sind die hessischen Bildungsstandards auf einen kumulativen Kompetenzerwerb von der Jahrgangsstufe 1 bis 10 ausgelegt, so dass den Lehrpersonen die Prinzipien des kompetenzorientierten Unterrichtens nahelegt wird (vgl. ebd., S. 7). Hierfür wurden in den Konzepten Erläuterungen zur Einbettung der Bildungsstandards in den bildungspolitischen Kontext sowie der Beitrag des jeweiligen Faches zur Bildung aufgenommen.

Neben inhaltlichen Fachaspekten wurde insbesondere die horizontale und vertikale Vernetzung von Lernprozessen fokussiert. Zusätzlich zu den Kompetenzbereichen des Unterrichtsfaches wurden überfachliche Kompetenzen, untergliedert in personale Kompetenz, soziale Kompetenz, Lern- und Arbeitskompetenz sowie Sprachkompetenz, formuliert, die für alle Fächer gleichermaßen verbindlich sind.

Die konkreten Bildungsstandards sind ebenfalls als Könnensleistungen ausgedrückt und als Regelstandards zu charakterisieren. Sie beziehen sich wie die nationalen Bildungsstandards auf das Ende der Primarstufe sowie auf die einzelnen Schulabschlüsse am Ende der Sekundarstufe I. In Form einer Synopse werden die Standards der Primarstufe und der Sekundarstufe I gegenübergestellt, so dass der angestrebte kumulative Kompetenzerwerb sichtbar wird. Zudem werden als Zwischenschritte in diesem Kompetenzerwerbsprozess lernzeitbezogene Kompetenzerwartungen und Inhaltsfelder für die Doppeljahrgangsstufen 5/6 und 7/8 ausgewiesen (vgl. ebd., S. 9 ff.).

Die Rohfassungen für die hessischen Bildungsstandards lagen im Dezember 2009 vor. Im folgenden Frühjahr wurden die Entwürfe im Internet veröffentlicht. Die Schulen, Studienseminare und Lehrerverbände erhielten damit die Möglichkeit, ein Feedback zu den Entwürfen zu formulieren und Entwicklungsmöglichkeiten aufzuzeigen. Die endgültigen Konzepte der hessischen Bildungsstandards wurden anschließend verabschiedet und sind mit Beginn des Schuljahres 2011/2012 in Kraft getreten. Damit sollten die bisherigen Lehrpläne für die Primarstufe und Sekundarstufe I abgelöst werden (vgl. ebd., S. 7).

Zur Konkretisierung der Standards und Inhaltsfeldern obliegt es den Schulen, ein eigenes Schulcurriculum zu entwickeln, dessen verbindliche Grundlage die Bildungsstandards sind. Im Schulcurriculum soll eine Zusammenführung der ausgewiesenen überfachlichen und

fachlichen Kompetenzen erfolgen, so dass konkrete Themen und Inhalte für die einzelnen Jahrgangsstufen festgelegt werden, anhand denen ein kompetenzorientierter Unterricht stattfinden kann (vgl. ebd., S. 11). Mit den Bildungsstandards und dem Schulcurriculum werden somit vermehrt Entscheidungsbefugnisse und Verantwortlichkeiten der Einzelschule übertragen, woraus eine größere curriculare Freiheit sowie mehr Selbstständigkeit und Eigenverantwortung für die Lehrkräfte resultieren. Hieraus ergeben sich neue Anforderungen an die Lehrerprofessionalität und die Arbeitsstrukturen innerhalb der Schule (vgl. ebd., S. 12). Die Implementierung der Bildungsstandards und Schulcurricula berührt somit alle drei Schulentwicklungsbereiche - die Unterrichts-, Organisations- und Personalentwicklung - in besonderem Maße. Dass die Schulen auf diese Herausforderungen vorbereitet werden müssen und die Initiierung solcher Entwicklungsprozesse Zeit und Unterstützung bedarf, erscheint selbstverständlich.

Auf die Ankündigung der Erstellung von Schulcurricula äußerten insbesondere die Lehrerverbände Kritik, da die Konzeption solcher Schulcurricula eine zu große personelle, organisatorische sowie zeitliche Belastung für die Schulen darstelle (vgl. Gewerkschaft Erziehung und Wissenschaft - Landesverband Hessen, 2010; Hessischer Philologenverband, 2010). Als Reaktion auf die Forderungen der Lehrerverbände hob das Hessische Kultusministerium die Verbindlichkeit zur Erstellung eines Schulcurriculums auf. Zukünftig ist dies keine verpflichtende Vorgabe, sondern die Schulen können nach Wunsch ein für ihre Arbeit grundlegendes Schulcurriculum entwickeln. Andernfalls gelten in Ergänzung zu den Bildungsstandards weiterhin die bisherigen Lehrpläne (vgl. Hessisches Kultusministerium, 10.02.2011, S. 1 f.). Hiermit ist jedoch die Gefahr verknüpft, dass die schulischen Akteure aus Gewohnheit und der derzeitigen Distanz zu den Bildungsstandards weiterhin lediglich mit den herkömmlichen Lehrplänen in der Unterrichtspraxis arbeiten, die keineswegs an Kompetenzen ausgerichtet sind. Dies würde die Akzeptanz und Integration der Standards in den schulischen Alltag erheblich verzögern und ein kompetenzorientiertes Unterrichten erschweren.

8.2 Das Konzept der Lernstandserhebungen in Hessen

Ein weiteres Aufgabenfeld des LSA beziehungsweise des vormaligen Instituts für Qualitätsentwicklung umfasst die Standardsicherung in Form der Erprobung, Durchführung und Organisation der Vergleichsarbeiten. In Hessen werden die Tests in der Sekundarstufe I als Lernstandserhebungen bezeichnet, da der Begriff „Vergleichsarbeit“ bereits im Rahmen von verbindlich zu schreibenden Parallelarbeiten in hessischen Schulen verwendet wird. Daher wurde die Titulierung „Lernstandserhebung“ ähnlich wie auch in anderen Bundesländern

nach der Übertragung der zentralen Testentwicklung an das IQB beibehalten. Um der hessischen Bezeichnung gerecht zu werden, wird im Folgenden sowie in der Auswertung der vorgenommenen Studie ebenfalls die Bezeichnung „Lernstandserhebung“ verwendet, während in Bezug auf Hessen mit dem Begriff „Vergleichsarbeit“ die erwähnte interne Parallelarbeit gemeint ist.

8.2.1 Implementierung und Teilnahme an den Lernstandserhebungen

Die Lernstandserhebungen wurden in Hessen schrittweise eingeführt und erprobt. Im Schuljahr 2006/2007 starteten die ersten Testläufe in der sechsten Klassenstufe in den Fächern Deutsch und Englisch sowie in der achten Klassenstufe in Mathematik (kurz: Lernstand 6 und Lernstand 8). Im folgenden Schuljahr kam die Leistungsmessung in der dritten Jahrgangsstufe in Deutsch und Mathematik (kurz: Lernstand 3) unter der damaligen Bezeichnung „Orientierungsarbeit“ hinzu. In diesen beiden ersten Testdurchläufen stand das Kennenlernen der Aufgabenformate und der didaktischen Materialien für die Nutzer in den Schulen im Vordergrund. Es erfolgte zu diesem Zeitpunkt noch keine Ergebnisrückmeldung an die Schulen. Seit dem Schuljahr 2008/2009 werden zusätzlich im Lernstand 6 Mathematik, im Lernstand 8 Deutsch sowie die erste Fremdsprache (Englisch, Französisch) getestet.
Zum gleichen Zeitpunkt wurde ein Lernstandsportal im Internet eröffnet, durch das die Lehrkräfte und Schulleitungen die notwendigen Informationen und Materialien erhalten sowie die Schülerergebnisse eingeben müssen. Sie bekommen zudem eine aufbereitete Rückmeldung zu ihrer Klasse zur Verfügung gestellt, die von der Universität Jena erstellt wird. Folglich stand im Testdurchlauf 2008/2009 insbesondere die Erprobung des Rückmeldeformats im Fokus (vgl. Institut für Qualitätsentwicklung, 2008, S. 5; Korngiebel, 2009, S. 40). Im zu betrachtenden Schuljahr 2009/2010 nahmen 18.660 Schüler am Lernstand 6 und 13.046 Schüler am Lernstand 8 teil (vgl. Institut für Qualitätsentwicklung, 2010, S. 1).
Die Einführung der Lernstandserhebungen erfolgte bereits vor der Implementierung der Bildungsstandards in Hessen. Damit ist die Überprüfungsfunktion der Tests, nach der die erreichten Kompetenzen bei den Schülern ermittelt werden sollen, sogleich hinfällig, da die Unterrichtsprozesse bislang noch nicht an den Standards ausgerichtet wurden. Vielmehr tritt in Hessen diesbezüglich die Innovationsfunktion der Tests in den Vordergrund, indem die Lehrkräfte über kompetenzorientiertes Testmaterial und didaktische Erläuterungen eine erste Annäherung an die Bildungsstandards erfahren. Dieser Aspekt ist insbesondere für die Auswertung der vorgenommenen Studie zu berücksichtigen, die zu dem Zeitpunkt durchge-

führt wurde, als die Entwürfe der hessischen Bildungsstandards wenige Wochen zuvor veröffentlicht wurden und die Lehrkräfte sich erstmalig mit ihnen auseinandersetzen konnten.
Während in der Primarstufe die Partizipation an den Lernstandserhebungen verpflichtend ist, wird in der Sekundarstufe I bislang ein Entwicklungszeitraum gewährt, in welchem schulische Strukturen und Erfahrungen zum Umgang mit dieser Art der Leistungsmessung aufgebaut werden sollen. Die Teilnahme am Lernstand 6 beruht auf dem Prinzip der Freiwilligkeit. Dabei entscheidet die Schulleitung selbst darüber, ob und in welchen Fächern Klassen an den Tests partizipieren. Folglich sind eine intrinsische Motivation und ein gewisses Interesse an den Testergebnissen die Voraussetzungen für die Entscheidung zur Teilnahme (vgl. Hessisches Kultusministerium, 2009, S. 27). Für den Lernstand 8 ist die Teilnahme in einem Unterrichtsfach verpflichtend, während die Anmeldung für die zwei weiteren Fächer wiederum freiwillig ist. Diese teilweisen Verpflichtung entstand aufgrund der Zielsetzung, dass die Möglichkeit schulinterner Ergebnisanalysen zwischen Parallelklassen und der Austausch von Fachkollegen erhalten bleibt (vgl. Institut für Qualitätsentwicklung, 2008, S. 5).

8.2.2 Kommunikation zwischen dem LSA und der Einzelschule

Für die Durchführung der endgültigen Tests wird an jeder teilnehmenden Schule ein Koordinator bestimmt. Während der Organisationsphase im Vorfeld der Lernstandserhebungen sowie im Zeitraum der Testdurchführung kommuniziert das LSA ausschließlich mit dem Schulkoordinator auf direktem Weg. Verbunden ist hiermit eine klare Aufgabenverteilung zwischen dem Institut, dem Koordinator und den teilnehmenden Lehrkräften. Der Koordinator leitet als Ansprechpartner die notwendigen Informationen zur praktischen Durchführung an die betreffenden Lehrkräfte und an die Schulleitung weiter und organisiert den formalen Ablauf der Tests in der Schule. Des Weiteren erhält er vom IQB konzipierte Handreichungen, Durchführungs- und Korrekturmanuale, nimmt die Testhefte an und prüft alle Materialien auf Vollständigkeit (vgl. Institut für Qualitätsentwicklung, 2008, S. 13). Der Koordinator stellt somit die einzige institutionalisierte Schnittstelle im Kommunikationskanal zwischen der Einzelschule und dem LSA dar. Zudem bietet das Institut eine Schulung der Koordinatoren im Vorfeld der Tests an. Erst während der Korrekturphase haben die Lehrkräfte die Möglichkeit, direkten Kontakt mit dem LSA aufzunehmen, indem sie sich bei Fragen oder Problemen an eine telefonische Hotline-Beratung wenden können.
Das Konzept der indirekten Kommunikation über den Schulkoordinator bietet Vorteile: Das LSA hat für jede Schule einen zentralen Ansprechpartner und kann die Informationen somit gebündelt weitergeben, was ein ressourcensparendes Vorgehen darstellt. Für die Lehr-

personen ist es möglicherweise eine Erleichterung, sich nicht zusätzlich mit unbekannten externen Personen in Verbindung setzen zu müssen, sondern alle notwendigen Materialien und Informationen durch eine Person ihres Kollegiums zu erhalten. Als Nachteil dieses Kommunikationsweges ist die Möglichkeit zu erwähnen, dass spezifische Informationen nicht korrekt oder nur unvollständig weitergegeben werden und somit Fehler bei der Durchführung oder der Korrektur entstehen können. Eine Form der Vergütung der Koordinationstätigkeit ist nicht vorgesehen.

Während des Nutzungsprozesses erhalten die Schulkollegien keine weitere persönliche Unterstützung durch das LSA. Es existieren Informationsblätter, welche Anregungen für die einzelnen Nutzungsphasen beinhalten. Diese bestehen aus standardisierten Formulierungen. Daher stehen die Lehrkräfte vor der Herausforderung, die Hinweise zunächst auf ihre individuelle Situation zu übertragen und anschließend spezifische Maßnahmen umzusetzen. Während der Konzeption von Handlungsstrategien und deren Realisierung ist keine Interaktion zwischen der Schule und dem LSA vorgesehen.

Die Lernstandserhebungen selbst werden vom LSA vorwiegend als ein Diagnoseinstrument zur Standortbestimmung, Stärken-Schwächen-Analyse und somit zur internen Evaluation verstanden und geworben. Schulbezogene Daten und Ergebnisse werden nicht veröffentlicht und alle Prozesse erfolgen anonymisiert. Dementsprechend findet kein Ranking statt und die Ergebnisse werden auch nicht zu einem Systemmonitoring genutzt. Des Weiteren dürfen die Lehrkräfte die Testergebnisse ihrer Schüler aus schulrechtlichen Gründen nicht benoten, da im Sinne eines kompetenzorientierten Lernprozesses die längerfristigen Unterrichtsprozesse bei den Lernstandserhebungen überprüft werden (vgl. Hessisches Kultusministerium, 2009, S. 10).

8.2.3 Rückmeldebausteine

Im Abschnitt 4.3.6 wurde detailliert erläutert, dass Qualität und Quantität des jeweiligen Rückmeldekonzepts die Nutzung der Testergebnisse massiv beeinflussen. Aus diesem Grund ist es sinnvoll, das in Hessen verwendete Rückmeldekonzept zu analysieren. Dem Projekt „kompetenztest.de“ des Lehrstuhls für Methodenlehre und Evaluationsforschung der Universität Jena wurde der Auftrag erteilt, die Auswertung und Aufbereitung der Rückmeldung zu den hessischen Lernstandserhebungen vorzunehmen (vgl. Hessisches Kultusministerium, 2009, S. 12). Unabhängig davon, ob es sich um den Lernstand 6 oder den Lernstand 8 handelt, weisen alle Rückmeldungen einen einheitlichen Aufbau auf. Sie

bestehen aus drei wesentlichen inhaltlichen Komponenten, welche die Lehrkräfte über das Lernstandsportal downloaden können.

- Tabellarische Auswertung,
- Sofortbericht,
- Ergebnisbericht.

Die *tabellarische Auswertung* kann als eine digitale Variante des Erhebungsbogens begriffen werden, den jede Lehrkraft bei der Korrektur ausfüllt. Auf diesem werden für die anonymisierten Schüler die erreichten Punkte jedes einzelnen Items aufgelistet. Daher stellt die tabellarische Auswertung für die Lehrperson zunächst keine neue Information dar. Allerdings werden zusätzlich die Punkte eines Schülers differenziert nach den Kompetenzbereichen und Anforderungsbereichen beziehungsweise GER-Niveaustufen aufaddiert und prozentual zur möglichen Gesamtpunktzahl ausgedrückt. Die tabellarische Auswertung erhält die Lehrkraft unmittelbar nach der Ergebniseingabe.
Zeitgleich kann der *Sofortbericht* eingesehen werden, in dem die Klassenergebnisse visualisiert darstellt werden. Zum einen wird in Form eines Balkendiagramms der erreichte Klassendurchschnitt prozentual zur möglichen Gesamtpunktzahl ausgedrückt. Dieser wird zusätzlich nach Kompetenzbereichen, Anforderungsbereichen beziehungsweise GER-Niveaustufen sowie nach Aufgaben differenziert abgebildet (vgl. Abbildung 19).
Zum anderen erhält die Lehrkraft eine tabellarische Zuordnung der einzelnen Items zu den Kompetenzbereichen und Anforderungsbereichen beziehungsweise GER-Niveaustufen sowie erste Impulse zur Auswertung dieser Diagramme mittels standardisierter Leitfragen. Offen bleibt zu diesem Zeitpunkt eine Bewertung der Ergebnisse. Was bedeutet es, wenn die Klasse beispielsweise durchschnittlich 70 Prozent der möglichen Punkte erreicht hat? Zudem wird offensichtlich, dass die Auswertung mit einfachen rechnerischen Methoden vorgenommen wurde. Die Punkte werden lediglich aufaddiert und prozentual ausgedrückt. Eine kriteriale Verortung zu den Kompetenzstufen findet nicht statt. Folglich stellen die Diagramme lediglich eine visuelle Darstellung gebündelter Ergebnisse dar, über welche die Lehrkraft bereits im Vorfeld durch die Korrektur verfügte.

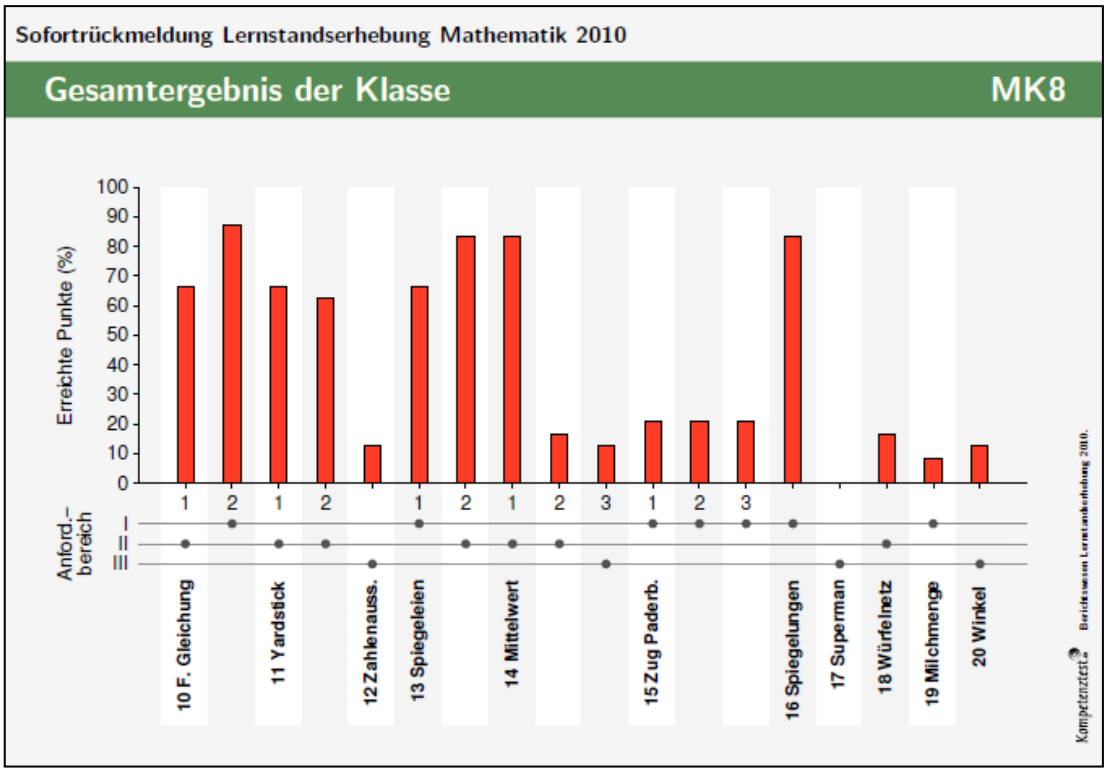

Abbildung 19: Beispielhaftes Diagramm des Sofortberichts - Durchschnittliches Gesamtergebnis der Klasse in den einzelnen Aufgaben (Institut für Qualitätsentwicklung, 2010, S. 3).

Zudem muss kritisch angemerkt werden, dass die Ergebnisse ausschließlich Durchschnittswerte einer gesamten Klasse wiederspiegeln. Eine Individualdiagnostik wird aus den im Abschnitt 4.3.5 dargelegten Begründungen nicht vorgenommen. Dies steht grundsätzlich dem Anspruch eines kompetenzorientierten Unterrichts nach Differenzierung entgegen. Die Diagramme verführen dazu, die Klasse als eine homogene Einheit zu begreifen, wodurch aus den Ergebnissen jedoch nur geringfügig diagnostische Erkenntnisse gewonnen werden können. Dieser Problematik wirkt ein weiteres Diagramm entgegen, welches die Punkteverteilung innerhalb der Klasse durch eine Quartilsbildung darstellt (vgl. Abbildung 20).

In diesem Diagramm werden jeweils die Schüler zusammengefasst, die das schwächste Viertel beziehungsweise das stärkste Viertel bilden. Die übrigen Schüler werden als Mittelgruppe begriffen. Mittels dieser Quartilierung wird eine Rangfolge gebildet, um das Abweichungsmaß der Ergebnisse zu erkennen. Die Beurteilung einer Schülerleistung im Vergleich zu den Mitschülern ist jedoch nicht möglich, da die Schüler als Punkte dargestellt werden und somit nur schwer zu identifizieren sind. Jedoch sagt auch dieses Diagramm nichts über die Wertigkeit der Leistungen aus, was als Nachteil der Quartilierung angeführt werden kann. Wenn beispielsweise nur wenige Schüler im höheren Punktebereich abgeschnitten haben, kann sich auch ein Schüler mit niedriger Punktzahl im obersten Quartil befinden.

Demgegenüber kann ein erfolgreicher Schüler in das unterste Quartil eingeordnet werden, wenn alle anderen Schüler gleichermaßen gut abgeschnitten haben, was eine negativere Bewertung seiner Leistung zur Folge haben kann (vgl. Institut für Qualitätsentwicklung, 2009, S. 10 f.). Das Ausmaß der Streuung der Ergebnisse ist somit die wesentliche Aussage dieses Diagramms.

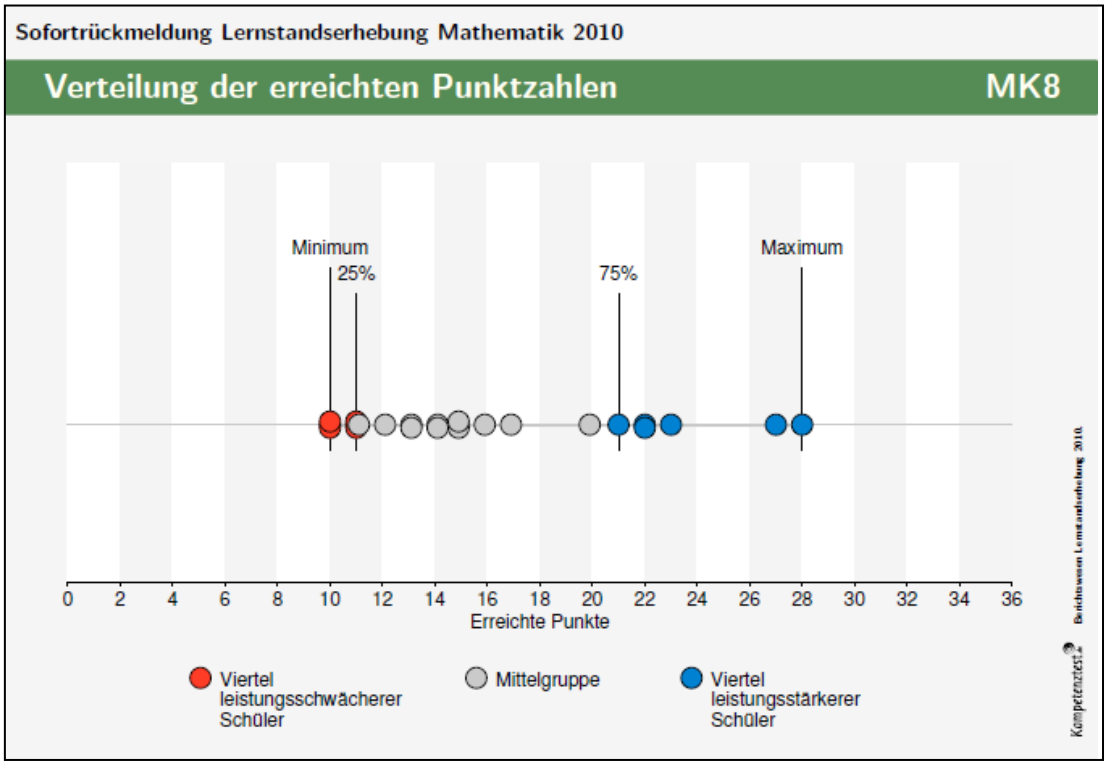

Abbildung 20: Beispielhaftes Diagramm des Sofortberichts - Verteilung der erreichten Punktzahlen innerhalb der Klasse (vgl. Institut für Qualitätsentwicklung, 2010, S. 3).

Zwölf Wochen nach der Ergebniseingabe wird den Lehrkräften der *Ergebnisbericht* zur Verfügung gestellt. Dieser umfasst grundsätzlich die gleichen Diagramme wie der Sofortbericht. Jedoch wird als zusätzliche Information im Rahmen des fairen Vergleichs jeweils der korrigierte Landesmittelwert als Referenzwert eingefügt. Zudem werden knappe Interpretationshinweise angeführt. Wenn beispielsweise die Streuung bei der Quartilierung verhältnismäßig groß ist, wird von einer heterogenen Lerngruppe ausgegangen und der Hinweis „Binnendifferenzierung“ in Bezug auf die Weiterarbeit geäußert. Zudem beinhaltet der Ergebnisbericht detailliertere Auswertungsimpulse sowie Informationen zum fairen Vergleich. Die bisherigen Erkenntnisse der Lehrpersonen werden somit im Ergebnisbericht um die Dimension des sozialen Vergleichs mit dem Landesmittelwert erweitert, wodurch eine Verortung der Leistungen ermöglicht wird. Eine tiefergehende qualitative und kriteriale

Auswertung in Bezug auf Kompetenzbereiche und -stufen findet nicht statt; es werden lediglich quantitative Werte ausgedrückt.

8.2.4 Didaktische Materialien

Neben den Rückmeldungen erhalten die Lehrkräfte didaktische Materialien zu den Testheften, die vom IQB in Berlin für den Lernstand 8 beziehungsweise von der länderübergreifenden Testentwicklergruppe für den Lernstand 6 erstellt werden. Die Materialien weisen kein einheitliches Design auf und haben verschiedene inhaltliche Schwerpunkte. In beiden Konzepten werden jeweils die Lernstandserhebungen in Hinblick auf die Testtheorie und ihren Bezug zu den Bildungsstandards erläutert. Teilweise werden zudem Erörterungen der Bildungsstandards im jeweiligen Fach sowie der theoretischen Grundlagen des kompetenzorientierten Unterrichts vorgenommen. Anschließend erfolgt eine Analyse jedes Testitems mitsamt seiner Lösung und den zu erwartenden Schwierigkeiten. Gegebenenfalls ist eine Kommentierung der zugrundeliegenden Anforderung des Items ebenfalls beinhaltet. Des Weiteren wird erneut eine tabellarische Zuordnung des Items zum Kompetenzbereich und Anforderungsbereich/ GER-Niveaustufe abgebildet. Eine Begründung dieser Zuordnung ist in den verschiedenen Materialien unterschiedlich ausführlich formuliert. Teilweise werden zu einzelnen Items Vorschläge für die Weiterarbeit im Unterricht geäußert.
In den Materialien zum Lernstand 8 in Deutsch und Mathematik wird zudem eine Schwerpunktsetzung vorgenommen. Dabei wird bei jedem Testdurchlauf ein zentraler Kompetenzbereich wie „Hörverstehen“ oder „mathematisches Argumentieren“ herausgegriffen und dessen theoretischen Grundlagen sowie Möglichkeiten zur Anbahnung des Kompetenzerwerbs im Unterricht detailliert und mit Beispielen versehen erläutert. Die didaktischen Materialien unterstützen auf diese Weise insbesondere die Innovationsfunktion der Lernstandserhebungen, indem die Bezüge zwischen den Testitems und den Bildungsstandards im Fokus stehen und den Lehrkräften das Konzept des kompetenzorientierten Unterrichts dargelegt wird.

8.2.5 Feedback-Erhebung

Zur Evaluation der hessischen Lernstandserhebungen werden die Lehrkräfte, Schulkoordinatoren und Schulleitungen darum gebeten, einen Online-Fragebogen zu beantworten. Auf Grundlage dieser Ergebnisse sollen die Organisation und die Durchführung der Tests verbessert sowie Informationen zur Nutzung der Tests gewonnen werden. Die schulischen Akteure können einen ersten Fragebogen unmittelbar nach Eingabe der Ergebnisse im Lernstandsportal beantworten. Die inhaltlichen Aspekte dieser Erhebung sind folgende:

- Bewertung der Materialien, der Ergebniseingabe, der organisatorischen Abwicklung sowie der Rückmeldebausteine,
- Analyse des Zeitmanagements in Bezug auf die Vorbereitung und der organisatorischen Abwicklung über den Koordinator,
- Erörterung festgestellter Probleme,
- Verwendbarkeit der Tests für die Diagnostik sowie für die Schul- und Unterrichtsentwicklung,
- Anregungen zu der qualitativen Verbesserung der Materialien, der Durchführung und der Nützlichkeit,
- Konsequenzen zur kooperativen Verwertbarkeit der Ergebnisse in verschiedenen Arbeitsstrukturen,
- Darlegung des ergriffenen Maßnahmenkatalogs.

Die Fragebögen für die Lehrkräfte, Schulkoordinatoren und Schulleiter unterscheiden sich inhaltlich voneinander, indem jeweils eine Fokussierung auf das jeweilige Tätigkeitsfeld der Person vorgenommen wurde. Beispielsweise beantworten die Koordinatoren lediglich Fragen zur Durchführung und zur Organisation der Tests.

Nach dem Erhalt des Ergebnisberichts werden die teilnehmenden Lehrpersonen gebeten, einen zweiten Fragebogen zu beantworten. Dieser greift nochmals eine Bewertung der Rückmeldebausteine und der Nützlichkeit der Tests sowie die Art der kollegialen Auswertung der Ergebnisse mitsamt den abgeleiteten Maßnahmen auf.

Die verschiedenen Fragebögen bestehen überwiegend aus geschlossenen Antwortformaten. Lediglich bei den Verbesserungsvorschlägen und der Aufzählung ergriffener Maßnahmen erhält der Beantwortende die Gelegenheit, sich ausführlicher zu äußern. Dieser Feedback-Bogen stellt im Regelfall die erste und einzige Möglichkeit für die Lehrkräfte dar, mit dem LSA zu kommunizieren (wenn auch auf einseitigem Weg) und ihre Einschätzungen und Erfahrungen auszudrücken (vgl. Abschnitt 8.2.2). Dabei bezieht sich der Feedback-Bogen

vorrangig auf die Organisation und die Gestaltung der Materialien und Rückmeldungen. Eine Auseinandersetzung mit den konkreten Testinhalten findet nicht statt.
Die Beantwortung der Fragebögen beruht auf Freiwilligkeit. Das hat zur Folge hat, dass eventuell ein nicht unbedeutender Anteil an Schulleitungsmitgliedern, teilnehmenden Lehrpersonen oder Koordinatoren an der Evaluation nicht teilnimmt und deren Erfahrungen nicht erfasst werden können. Des Weiteren bleibt es für die Lehrpersonen unklar, wozu die Ergebnisse des Fragebogens verwendet werden, da eine Veröffentlichung der Auswertung zumindest zum Lernstand 6 und Lernstand 8 bislang unterblieb. Lediglich für den Lernstand 3 wurden einzelne Informationen bekannt gegeben. Demnach wurden die Materialien und Rückmeldebausteine von den schulischen Akteuren durchgängig als befriedigend bewertet. Zudem nahmen 83,3 Prozent der Lehrkräfte den Sofortbericht und 61,4 Prozent der Lehrkräfte den Ergebnisbericht zur Kenntnis. Auf die didaktischen Materialien griff hingegen nur jeder Fünfte zu (vgl. Institut für Qualitätsentwicklung, 2010, S. 32).
Das LSA verwendet die Erkenntnisse aus der Evaluation für die Verbesserung der organisatorischen Abläufe, wie am Beispiel des Lernstands 3 ersichtlich wird. Nachdem Lehrkräfte und Lehrerverbände öffentlich zu der Organisation, dem Inhalt und der Korrektur der Tests massive Kritik äußerten, übernahm das Institut ab dem Schuljahr 2010/2011 die Druck- und Versandkosten, veröffentlichte im Vorfeld die zu testenden Kompetenzbereiche und splittete den Mathematiktest auf zwei Testzeitpunkte auf (vgl. ebd., S. 28).

8.3 Weitere Instrumente zur standardisierten Leistungsmessung in Hessen

Im Rahmen der standardisierten Qualitätssicherung nimmt Hessen neben den Lernstandserhebungen an den in Abschnitt 3.2.1 erwähnten internationalen und nationalen Erhebungen teil. Seit 2004 finden zudem zentrale Abschlussprüfungen für den Hauptschulabschluss und mittleren Schulabschluss sowie seit 2007 für das Abitur statt. Des Weiteren können Verfahren zur Prüfung des Schriftspracherwerbs und des Leseverstehens freiwillig in den Klassenstufen 2, 7 und 9 genutzt werden (vgl. Hessisches Kultusministerium, 2009, S. 8 f.).
Im Fach Mathematik existiert zusätzlich zum Lernstand 8 der Mathematikwettbewerb in der achten Jahrgangsstufe, der bereits 1969 in Hessen eingeführt wurde. Hierbei werden in drei Durchführungsrunden auf Schul-, Kreis- und Landesebene die Leistungen der Schüler in Form eines Wettbewerbs flächendeckend getestet. Eine Differenzierung erfolgt nach Schulformen. Die Lehrkräfte korrigieren die Arbeiten selbst und können diese als Klassenarbeit werten. Eine Rückmeldung erhalten sie nicht. Das durchschnittliche Schulergebnis wird an das Hessische Kultusministerium weitergeleitet und in Form eines Leistungsvergleichs mit-

tels einer Quartilierung veröffentlicht. Der Mathematikwettbewerb ist bislang inhaltlich nicht an den Bildungsstandards ausgerichtet.

Die „hessischen Vergleichsarbeiten“ (nicht zu verwechseln mit den als „Lernstandserhebungen“ implementierten bundesweiten Vergleichsarbeiten) wurden im Schuljahr 2001/2002 als verbindlich zu schreibende Parallelarbeiten eines Jahrgangs in den Klassenstufen 6 und 8 in den Hauptfächern eingeführt. Sie sind somit kein externes Testinstrument. Mithilfe der Parallelarbeiten sollen die Lehrkräfte zu einem kollegialen Austausch und zu kooperativer Unterrichtsentwicklung angeregt werden. Die Lernstandserhebungen ersetzen diese Arbeiten nicht. Aus diesem Grund werden die „hessischen Vergleichsarbeiten“ zusätzlich zu den Lernstandserhebungen und teilweise in den gleichen Jahrgangsstufen der Sekundarstufe I durchgeführt (vgl. Korngiebel, 2009, S. 39).

9 Nutzung der Lernstandserhebungen entsprechend des Zyklenmodells nach Helmke

Die Darstellung der Untersuchungsergebnisse der qualitativen Interviewstudie an zwölf hessischen Gymnasien orientiert sich am Zyklenmodell von Helmke (vgl. Helmke A. , 2004, S. 100) (vgl. Abschnitt 5.5). Dementsprechend wird zunächst die Nutzung der Lernstandserhebungen durch die schulischen Akteure in den untersuchten Schulen differenziert nach den einzelnen Phasen des Nutzungsprozesses, der Rezeption, Reflexion, Aktion und Evaluation, beleuchtet. Anschließend erfolgt eine Auswertung, inwiefern weitere Faktoren (individuelle, schulische und externe Bedingungen) den Nutzungsprozess beeinflusst haben.
Zur Untermauerung der Aussagen werden Interviewausschnitte zitiert. Die Zuordnung des Zitats zu dem zugehörigen Probanden ist zu erkennen, indem beispielsweise die Angabe „LK 1“ die Lehrkraft der Schule 1 und dementsprechend „SL 6“ das Schulleitungsmitglied der Schule 6 meint. Eine Übersicht über die Probanden ist in Abschnitt 6.3.1 einzusehen. Die vollständigen Transkriptionen der Interviews und der Nachgespräche sind im beiliegenden Anhang aufgeführt. Um eine Verortung des Zitats im Gesamtinterview zu erleichtern, erfolgt am Ende des Ausschnitts zudem die Angabe des Zeitpunkts im Interview.

9.1 Nutzungsphase der Rezeption

Entsprechend dem Zyklenmodell von Helmke (vgl. Helmke A. , 2004, S. 100) beginnt nach Erhalt der Ergebnisrückmeldung der Nutzungsprozess bei den Lehrkräften und Schulleitungen. Die erste Phase bildet hierbei die Rezeption, in welcher die Rückmeldung selbst sowie die zur Verfügung stehenden Materialien angenommen und betrachtet werden. Dies bildet die Voraussetzung für die sich anschließende Analyse der Testergebnisse. Aus diesem Grund ist die Untersuchung bedeutsam, in welcher Intensität die Lernstandserhebungen und Rückmeldungen von den teilnehmenden Lehrpersonen rezipiert wurden und inwiefern sich diese Nutzungsphase auf die Prozesse der Reflexion und Aktion auswirkte. Daher wurden die Aussagen der Interviewpartner zunächst auf den Aspekt „Rezeption“ bezogen analysiert.

9.1.1 Rezeptionsaspekte

Es bildeten sich bei der Untersuchung der Interviewaussagen in Bezug auf die Rezeption drei Bereiche heraus, die von den befragten Lehrpersonen und Vertretern der Schulleitungen in unterschiedlicher Intensität rezipiert wurden: Testinhalte, Rückmeldungskonzept sowie Handreichungen mit dem zugehörigem didaktischen Material.

Testinhalte

Die Rezeption der Testinhalte umfasste vorrangig ein Hinterfragen und Überprüfen des Testmaterials, wie aus den folgenden Interviewzitaten ersichtlich wird:

> *LK 2: Ja, ich will mal so sagen, ich bin die Aufgaben durchgegangen […] und habe versucht, die Aufgaben und ihre Zielsetzungen zu erschließen aufgrund der Korrekturanweisungen. (LK 2, 00:09:13-6)*
>
> *LK 2: Ja, ich bin dann von meinen Erfahrungen als Lehrer ausgegangen und habe da so ein oft vorbewusstes Repertoire von Themen und von Aufgaben, die man im Unterricht macht und davon ausgehend guckt man sich jetzt diese Aufgaben an. (LK 2, 00:21:04-3)*
>
> *SL 6: Ja, was mich mal interessiert, diese Aufgabenformate: Wo kommen die denn her oder aus welchen Quellen speisen sich denn so diese Aufgaben? Kommen die aus dem angelsächsischen, angloamerikanischen Unterrichtsbereich oder sind da deutsche Didaktiker am Werk? Woher kommt das? (SL 6, 00:21:04-3)*

Aus den Beispielen geht hervor, dass zunächst einmal wahrgenommen wurde, welche Kompetenzbereiche überhaupt getestet werden. Anschließend wurden einzelne Items näher betrachtet. Zentrale Kriterien bildeten hierbei die Schwierigkeit sowie die Lehrplanadäquatheit der Aufgaben. Für die teilnehmenden Lehrkräfte war es von besonderem Interesse zu untersuchen, inwiefern die Items den Anforderungen entsprechen, die sie selbst im Unterricht an ihre Schüler stellen beziehungsweise inwieweit die Aufgaben generell die zu unterrichtenden Themen abdecken.

Ein weiterer Aspekt der Rezeption stellte die Art und Weise des Testens, speziell die Aufgabenformate, dar. Wie aus der obigen Aussage von dem Schulleitungsmitglied 6 hervorgeht, weckte dies auch Interesse bezüglich der Hintergründe zur Konzeption und Entwicklung der Tests, was als unbewertete Neugierde zu verstehen ist. Ähnliche Anmerkungen traf auch eine weitere Lehrkraft.

Die Betrachtung der Testinhalte erfordert nicht den Erhalt der Rückmeldung. Daher setzt die Rezeptionsphase teilweise bereits unmittelbar nach der Durchführung der Tests ein und kann der Rezeption der Rückmeldung zeitlich vorgelagert sein. Jedoch muss an dieser Stelle darauf hingewiesen werden, dass lediglich vier der an den Test teilnehmenden Lehrpersonen die Testinhalte als einen Rezeptionsaspekt wahrnahmen. Folglich ist anzunehmen, dass

der Fokus während der Rezeption überwiegend auf den Inhalten der Rückmeldebausteine lag.

Rückmeldung

Bezüglich der Rückmeldung war bei den teilnehmenden Lehrkräften ein generelles Interesse an den Ergebnissen ihrer Schüler zu konstatieren, so dass die Berichte fast von der Gesamtheit der befragten Personen zumindest betrachtet wurden. Lediglich zwei Probanden erhielten die Rückmeldung nur teilweise oder gar nicht.

Eine Lehrkraft konnte ausschließlich den Sofortbericht und die tabellarische Auswertung rezipieren, da ihr der Ergebnisbericht nicht zur Verfügung gestellt wurde. Der Grund dafür war eine schulexterne Person, die im Rahmen einer geringfügigen Beschäftigung die geschlossenen Items der Schülerarbeiten korrigiert und die Ergebnisse eingegeben hatte. Als der Ergebnisbericht erschien, war diese Person, welche über das Passwort zum Lernstandsportal verfügte, nicht mehr an der Schule beschäftigt. Daher wurden die Ergebnisrückmeldungen an die teilnehmenden Lehrpersonen nicht weitergeleitet. Allerdings hätte die Lehrkraft auch selbst das Passwort beantragen und daraufhin eigentätig Zugang zum Lernstandsportal erhalten können. Es muss geschlussfolgert werden, dass das Interesse an dem Ergebnisbericht nicht groß genug gewesen war, um dies zu veranlassen. Allerdings ist hierbei auch die Organisation der Tests innerhalb dieser Schule als nicht zufriedenstellend zu beurteilen, da zumindest der Koordinator nach Ausscheiden der externen Person für eine Weitergabe der Rückmeldungen verantwortlich gewesen wäre.

Bei der anderen Lehrperson verhinderte eine Erkrankung die termingerechte Eingabe und somit den Erhalt von Rückmeldeberichten:

> *LK 10: Dazu muss ich sagen, dass ich diese Rückmeldung nicht bekommen habe aus einem bestimmten Grund. Und zwar gab es einen bestimmten Termin, bis zu dem die Daten von mir oder von allen eingegeben werden sollten, und genau in diese Phase fiel bei mir eine private Operation und ich konnte sozusagen den Termin nicht wahrnehmen und ich wollte dann 48 Stunden später noch eingeben, was mir aber […] dann nicht mehr möglich wurde. Die sagten halt, das Ganze ist schon abgeschlossen und ich konnte halt zwei Tage später nicht mehr eingeben. (LK 10, 00:04:22-3)*

Die Lehrkraft 10 bedauerte im weiteren Gespräch, dass das zum damaligen Zeitpunkt zuständige Institut für Qualitätsentwicklung nicht flexibel auf dieses Problem hat reagieren können, denn sie wäre sehr an Referenzwerten zu den Ergebnissen ihrer Klasse interessiert gewesen. Vorweggreifend soll an dieser Stelle angemerkt werden, dass die Lehrkraft 10 dennoch eine intensive Reflexion der Ergebnisse anhand ihrer Korrektur vorgenommen hat, so dass das Nichtvorhandensein der Rückmeldung den Nutzungsprozess in diesem Fall nicht

massiv behindert hat, während bei der zuvor betrachteten Lehrperson nur eine geringfügige weitere Nutzung zu konstatieren war. Dies lässt eine immense Einflussnahme der intrinsischen Motivation zur Ergebnisanalyse auf den Nutzungsprozess vermuten (vgl. Abschnitt 10.1.1).

Bis auf die Lehrkraft 10 ist bei allen befragten Lehrkräften und Vertretern der Schulleitungen, die selbst an den Lernstandserhebungen teilnahmen, eine Rezeption der einzelnen Rückmeldebausteine festzustellen. Dies impliziert jedoch keine Selbstverständlichkeit für die Rezeption, wie folgende Aussagen aufzeigen:

> *SL 4: Die Kollegen sind natürlich gebeten worden, sich das abzuholen da aus dem Internet, das auch zu vergleichen. Der eine oder andere macht das, aber ich muss ehrlich sagen, ich habe eine ziemlich große Lustlosigkeit unter den Kolleginnen und Kollegen verspürt, sich dann da nochmal mit zu beschäftigen. Die haben wirklich geflucht wie die Raben. (SL 4, 00:12:28-6)*

> *SL 6: Die Kollegen selbst machen das teilweise gar nicht. Also ich lege ihnen immer nahe: "Macht doch mal zunächst schon mal diese Sofortrückmeldung!" Aber gut, das kriegen dann einige schon nicht gebacken und dann weise ich sie darauf hin: "Macht es doch mal!" […] Diesen Termin, dass das dann quasi das Ganze mit diesem korrigierten Mittelwert vergleichbar ja ist, der dann landesweit erhoben ist, das kriegen die meisten dann schon nicht mehr mit oder wissen die Passwörter nicht oder es wird irgendwie so ein bisschen, naja, nicht mehr so verfolgt. (SL 6, 00:04:37-0)*

Ein nicht zu unterschätzender Anteil der Lehrkräfte scheint somit keine Rezeption der Rückmeldung vorzunehmen, was die weiterführende Nutzung selbstverständlich massiv beeinflusst und erschwert. Als Gründe wurden fehlende technische Fertigkeiten und vor allem mangelndes Interesse und zu geringe Motivation zur nochmaligen Beschäftigung mit den Tests angeführt. Insbesondere aus dem Zitat von Schulleitungsmitglied 4 kann geschlussfolgert werden, dass sich die Rezeptionsintensität des Sofortberichts auf die Rezeptionsintensität des Ergebnisberichts auswirkt. Wenn der Sofortbericht bereits als nicht besonders interessant eingeschätzt wurde und nur wenige Informationen daraus gewonnen werden konnten, ebbt das Interesse an einer Rezeption des Ergebnisberichts automatisch ab.

Dass in der Interviewstudie keine Personen angetroffen wurden, die keine Rückmeldungsrezeption durchgeführt haben, ist mit der Freiwilligkeit der Untersuchung zu erklären. Es nahmen ausschließlich Lehrkräfte und Schulleitungsvertreter teil, die Interesse an der Befragung hatten. Eine Person, die keine Rezeption durchgeführt hat, hätte wahrscheinlich auch kein Interesse an einer Beteiligung an einer Forschungsstudie gehabt. Umso aufschlussreicher sind daher die zwei dargestellten Äußerungen der Schulleitungsvertreter bezüglich der Rezeptionsintensität an ihrer Schule zu bewerten.

Generell konnte bei sechs von 19 teilnehmenden Lehrkräften ein geringfügiges Interesse an den Ergebnissen konstatiert werden, wie folgende Aussagen bestätigen:

LK 5: Ja (gedehnt), fand ich interessant, hatte ich aber ehrlich gesagt, hab ich mir dann gar nicht viel Zeit genommen, um da im Einzelnen zu gucken, wo stehen wir? Also ich fand es ok wahrzunehmen, wir waren im Mittelfeld da und das hat mir dann gelangt (lacht). (LK 5, 00:04:13-5)

SL 4: Wir können uns dann die Berichte anschauen und können sagen, gut, die haben alle […] durchschnittlich abgeschnitten […]. Also wenn man das vergleicht, es ist immer fast Landesdurchschnitt, eher ein bisschen darüber als ein bisschen darunter, aber das war es. (SL 4, 00:08:50-8)

SL 6: Die gezielte Auswertung, beispielsweise etwa was die Rangfolge innerhalb der Schule angeht oder was die Relation zu dem korrigierten Mittelwert angeht, die haben wir eigentlich eher noch ein bisschen beiseitegelassen. (SL 6, 00:03:07-6)

Bei dieser Form der Rezeption erfolgte vorrangig eine oberflächliche Betrachtung des Gesamtergebnisses im Vergleich zu den Landesmittelwerten. Das Interesse an den Ergebnissen beschränkte sich daher auf den sozialen Vergleich, um einen Eindruck zu erhalten, wie die eigene Klasse im Vergleich zu Referenzwerten positioniert ist. Tiefergehende Informationen wurden den einzelnen Rückmeldebestandteilen dabei nicht entnommen. Als Gründe wurden mangelndes Interesse, fehlende Zeit beziehungsweise der momentan geringe Stellenwert der Lernstanderhebungen im Kollegium und im Schulalltag angeführt.

Der soziale Vergleich mit den Landesmittelwerten stand jedoch nicht nur bei dieser Nutzungsform im Zentrum der Rezeptionstätigkeit. Neun von 19 teilnehmenden Lehrkräften führten an, dass der Mittelwert ihrer Klasse in Zusammenhang mit dem fairen Vergleich für sie die interessanteste Information der Rückmeldung darstelle.

LK 1: Also, ich habe mir das jetzt nochmal angeschaut auf den Mittelwert, also natürlich wie steht meine Klasse jetzt im Vergleich zum Landesdurchschnitt. Das ist ja schon interessant. (LK 1, 00:09:12-9)

LK 2: Die im nunmehr vorgelegten Ergebnisbericht enthaltenen Einzeldaten und Gesamtergebnisse werden in Säulendiagrammen vorgestellt und mit den entsprechenden Ergebnissen des Landesdurchschnitts verglichen. Ich halte diese Rückmeldung für wichtig und aufschlussreich.(LK 2, Bericht)

SL 12: Aber, sagen wir mal, im Vergleich zu anderen oder wie man so situiert ist in der Schule, spielt der ja schon irgendwo eine gewisse Rolle. Liegen wir darüber, liegen wir darunter? (LK 12 + SL 12, 00:10:15-3)

Die Bedeutsamkeit des sozialen Vergleichs wird hier signifikant deutlich. Die wesentliche Frage bei der Betrachtung ist daher, ob die Klasse über oder unter dem Landesmittelwert liegt. Lediglich eine Lehrperson fügte an, dass für sie der reine Vergleich des Klassenergebnisses mit den Referenzwerten unerheblich sei:

LK 12: Also der korrigierte Landesmittelwert hat mich eigentlich gar nicht interessiert, also dieser Durchschnittswert sage ich jetzt mal. Wenn das insgesamt, was

weiß ich, 10,39 war und wir waren bei 10,62 oder so, das war für mich völlig uninteressant. […] So ein Durchschnittswert, ok, den braucht man für die Statistik, aber den fand ich relativ unerheblich. (LK 12 + SL 12, 00:10:05-0)

In der Tat bietet der soziale Vergleich nur geringfügige Informationen, wenn der Abstand nicht besonders deutlich hervortritt. Hierzu muss berücksichtigt werden, dass bei diesem Diagramm der Rückmeldung lediglich das durchschnittliche Gesamtergebnis der Klasse dargestellt wird und noch keine spezifischeren Informationen enthalten sind.

Es ist anzunehmen, dass der Aufbau der Ergebnisrückmeldung die Fokussierung auf die Positionierung der Lerngruppe bei einem Großteil der Befragten verstärkt hat. Gleich zu Beginn bekommen die Lehrkräfte den Vergleich des Gesamtdurchschnittes mit dem Landesmittelwert zur Verfügung gestellt. Mit dieser Aussage ist ein erstes Informationsbedürfnis gestillt. Entweder liegen keine weiteren Informationswünsche vor und die übrigen Diagramme werden gar nicht beziehungsweise nur oberflächlich betrachtet, wie es zuvor beschrieben wurde, oder die betrachtende Person möchte weiterführende und detailliertere Erkenntnisse gewinnen und rezipiert den restlichen Teil des Ergebnisberichts. Im letzteren Fall wurden differenzierte Aussagen betrachtet, wie die Untergliederung nach Kompetenzbereichen oder aufgabenbezogene Auswertungen.

LK 4: Da haben wir die verschiedenen Kompetenzen bezüglich Schreiben, Hören und da ist zu sehen, bei welchen Aufgabentypen haben sie besonders gut oder besonders schwach abgeschnitten. […] Das fand ich für mich jetzt eine sehr interessante Sache. (LK 4, 00:05:14-1)

LK 12: Also ich habe vor allem geschaut: Aufgabe 1, wie schaut es denn da aus? Bei den einzelnen Aufgaben, wie liegen wir da? […] Also mich hat wirklich interessiert die Aufgabenbereiche, die einzelnen Kompetenzen und dann natürlich auch das Ergebnis bei dem einzelnen Schüler. (LK 12 + SL 12, 00:10:05-0)

Eine solche weiterführende Rezeption konnte insgesamt bei acht teilnehmenden Lehrkräften festgestellt werden. Interessanterweise stand auch bei der Betrachtung der differenzierteren Diagramme der soziale Vergleich für die jeweilige Lehrkraft im Vordergrund des Informationsbedürfnisses. Die Ergebnisdarstellung der Streuung der Klassenergebnisse mittels der Quartilierung wurde hingegen von nur zwei Befragten als Rezeptionsgegenstand erwähnt. Folglich bietet dieses Diagramm anscheinend lediglich geringfügig Informationen, so dass diese Erkenntnis als nicht besonders relevant eingestuft wird.

Generell wurden mit Ausnahme eines Probanden die Rückmeldebestandteile des Sofort- und Ergebnisberichts grundsätzlich als verständlich und aufschlussreich eingeschätzt. Die Rezeption erfolgte größtenteils auf die Klasse bezogen, wie es auch in den einzelnen Rückmeldebausteinen offeriert wird. Lediglich drei Personen versuchten, eine individuelle Auswertung der einzelnen Schülerleistungen vorzunehmen, um damit Informationen zu erhalten, die sie an die Schüler weitergeben konnten.

SL 2: Sie gibt Rückmeldungen, detaillierte Rückmeldungen über individuelle Fähigkeiten und individuelle Entwicklungsmöglichkeiten. Wo ist ein Schüler schon recht stark, wo zeigt er dies auch dann im Test? Und womit kommt er gar nicht so gut zurecht? (SL 2, 00:08:35-1)

Bei dieser Aussage muss jedoch berücksichtigt werden, dass das Schulleitungsmitglied 2 selbst bisher nicht an den Lernstandserhebungen teilgenommen hat und sich noch nicht intensiv mit den Testinhalten auseinandergesetzt hatte, die Tests jedoch generell als sehr bedeutsam für die Schulentwicklung einstufte. Aufgrund der fehlenden Erfahrung spiegelt dieses Zitat auch eine gewisse idealisierende Einschätzung der Tests wieder.

Zwar haben zwei weitere Lehrkräfte versucht, ebenfalls schülerspezifische Informationen mittels der tabellarischen Auswertung zu erhalten, doch von den restlichen Befragten wurde eine Rezeption der individuellen Ergebnisse nicht bestätigt. Ein möglicher Grund ist die Gestaltung der tabellarischen Auswertung, wie eine Lehrkraft anführte. Diese sei aufgrund ihres Formats unübersichtlich und schwer zu handhaben. Zudem ist es denkbar, dass die Lehrkräfte die tabellarische Auswertung lediglich als die digitale Version ihres Erhebungsbogens betrachtet haben und Erläuterungen zu den zusammengetragenen Punktzahlen einzelner Schüler in den jeweiligen Kompetenzbereichen und Anforderungsstufen benötigt würden. Für die technische Benutzung des Dokuments gibt es eine Handreichung. Diese ist den Lehrkräften jedoch nicht bekannt beziehungsweise wurde bei der Rezeption nicht als Hilfsmittel verwendet.

Resümierend kann konstatiert werden, dass bei allen Befragten die Betrachtung des sozialen Vergleichs in der Rezeption dominierte, währenddessen die Streuung der klasseninternen Ergebnisse als relativ unerheblich eingeschätzt wurde. Als erkenntnisreich wurde zudem die aufgabenbezogene Auswertung angeführt, die jedoch nachweislich von nur acht von 19 Lehrkräften vorgenommen wurde. Dieser Rezeptionsgegenstand kann jedoch als besonders gewinnbringend beurteilt werden, da er sowohl die Positionierung der Lerngruppe anhand des sozialen Vergleichs als auch eine Auseinandersetzung mit den Testinhalten fördert. Um das Ergebnis der Klasse zu einem Item angemessen beurteilen zu können, ist es notwendig, die Aufgabe mit ihren zugrundeliegenden Anforderungen und Zielsetzungen selbst zu betrachten. Demnach gab es verschiedene Bewertungen der einzelnen Diagramme und die jeweiligen Darstellungen und Informationen schienen das Interesse der Lehrkräfte nicht gleichermaßen angesprochen zu haben.

Insbesondere bei den Schulleitungsvertretern wurde ein massives Interesse am sozialen Vergleich als die wichtigste Information deutlich, so dass die reine Verortung der Klassen- und Schulergebnisse vorrangig forciert wurde. Eine differenziertere Rezeption erfolgte nur selten.

Der Erhalt der Rückmeldung führte nicht automatisch und nicht gleichermaßen zu der Rezeption ihrer Inhalte. Vielmehr sind immense Unterschiede in der Intensität der Rückmeldungsrezeption festzustellen. Als zentrale Einflussfaktoren hierfür sind das individuelle Interesse und die vorhandene Motivation anzuführen. Aber auch existierende Erfahrungen und bewertende Einschätzungen zu den Tests wirkten massiv auf die Betrachtung der Ergebnisse ein.

Handreichungen und didaktisches Material

Bezüglich der Rezeption der Handreichungen und der didaktischen Materialien kann festgestellt werden, dass die Anweisungen zur Durchführung und Korrektur durchgängig benutzt wurden:

> *LK 5: Also dieses Organisatorische war schon sehr hilfreich für mich zu gucken, ok, wie muss ich es konkret machen? Das war notwendig eigentlich, ja, um da für mich vorher die Materialien bereit zu haben, zu wissen, aha, wie mache ich das? Wann ist die Pause? Oder wie viel Zeit gebe ich da? Das war völlig in Ordnung. (LK 5, 00:25:03-3)*
>
> *LK 7: Ich habe mir das einfach durchgelesen. Da stand ja dann auch, wie man einfach das durchführen soll. Und das sind ja schon sehr strenge Regeln, dass man bei der CD nicht auf Pause drücken darf. Also man muss das schon haben, um das dann korrekt durchführen zu können, um einfach auch wieder die Vergleichbarkeit und Objektivität gewährleisten zu können. […] Also so etwas muss ja schon dabei sein. (LK 7, 00:37:47-9)*
>
> *SL 2: Insofern wurden die schon als sehr positiv und hilfreich erlebt, weil die vorher von allen Lehrern auch durchgelesen und durchgeguckt wurden, zur Vorbereitung der Tests. … [Es] ist nicht darüber gemeckert worden, das ist schon mal gut, dann muss es gut gewesen sein. Denn man hätte ja auch schnell eine Antwort bekommen: "Da produzieren die einen Haufen Papier und da steht nichts drin." Das ist nicht gekommen, also muss es wohl hilfreich sein. (SL 2, 00:48:08-3)*

Dementsprechend wurden die Vorgaben in den Handreichungen als verbindlich betrachtet und als Handlungsgrundlage für die Durchführung und die spätere Korrektur verwendet. Mit dem Bedürfnis, die Lernstandserhebungen den Vorschriften gemäß vorzunehmen, ist automatisch eine intensive Rezeption der Anleitungen verknüpft. Lediglich ein Proband führte an, dass diese Durchführungsmanuale nicht notwendig seien, da die Tests in der Weise selbsterklärend seien.

Als Kritik an den Handreichungen wurde der immense Umfang der Materialien angemerkt, mit dem ein nicht unerheblicher Arbeitsaufwand verknüpft sei. Hieraus kann geschlussfolgert werden, dass die Vorbereitung zu viel Zeit in Anspruch nimmt und die Anleitung knapper gestaltet werden könnte. Des Weiteren konnten Diskrepanzen zwischen den Korrektur-

vorgaben und den eigenen Bewertungsvorstellungen festgestellt werden. Dies wirkte sich zugleich auf die generelle Bewertung der Tests aus, wie im Abschnitt 11.3 dargestellt wird. Entgegen der relativ durchgängig intensiven Reflexion der Handreichungen erfolgt eine Betrachtung des zusätzlichen didaktischen Materials nur selten, wie folgende Zitate wiedergeben:

LK 10: Erachte ich für sehr sinnvoll, auch wenn ich sie nicht so genutzt habe.
I: Und warum haben Sie das nicht so genutzt?
LK 10: Aus Zeitgründen, weil das so in das schriftliche Abitur fiel, was ich halt auch betreut habe, und so weiter und so fort. Aber generell ist es natürlich sinnvoll. (LK 10, 00:25:31-9)

SL 11: Ich mache da, wie Herr [LK 11] auch, ich mache da irgendwann mal Cut. Ich hab das Ding durchgezogen und wenn da noch 5 MB Informationen auf der Homepage […] stehen, das ist mir dann egal. Die Zeit ist abgelaufen und das Abitur steht an oder sonst irgendetwas. Man kann sich doch nicht ohne Ende mit diesen Sachen befassen, weil man einfach die Zeit nicht hat! (LK 11 + SL 11, 00:53:51-4)

Dementsprechend wurden die didaktischen Materialien oftmals noch nicht einmal betrachtet. Als Hauptgrund für die fehlende Rezeption diente oftmals die fehlende Zeit. Jedoch stellten auch mangelhaftes Interesse und nichtvorhandene Motivation bedeutsame Faktoren dar. Es kann vermutet werden, dass die Lehrkräfte sich nicht über den Inhalt des didaktischen Materials bewusst waren beziehungsweise darin keinen Anknüpfungspunkt für die Verbesserung und Weiterentwicklung ihres eigenen Unterrichts sahen. Lediglich drei Lehrkräfte äußerten, dass sie sich das Material angesehen hätten, wobei die Standardisierung des Materials bei einer Lehrperson kritisch bewertet wurde. Die enthaltenen Informationen könnten nicht einfach auf die eigene Lerngruppe für die weitere Nutzung übertragen werden. Die allgemein gehaltenen Angaben und Hinweise seien in diesem Fall hinderlich, da zunächst eine Anpassung an die eigene Lernsituation notwendig sei. Aber dieses Problem besteht bei allen didaktischen Materialien, die den Lehrkräften - unabhängig von den Lernstandserhebungen - von den Schulbuchverlagen und sonstigen Anbietern zur Verfügung stehen. Daher kann die Angabe der Standardisierung als Hinderungseffekt kaum akzeptiert werden. Vielmehr schien die Motivation zu einer Beschäftigung mit dem Zielen und Merkmalen des kompetenzorientierten Unterrichts in Verbindung mit der Lernstandserhebung noch nicht ausreichend groß genug zu sein. Möglich ist auch, dass die Qualität der didaktischen Materialien nicht als ausreichend positiv eingeschätzt wurde, als dass sich nach dem Empfinden der jeweiligen Lehrkraft der Aufwand für eine intensivere Rezeption gelohnt hätte.

LK 12: Also das habe ich überflogen und habe gedacht: Weg! (lacht) (LK 12 + SL 12, 00:52:49-8)

SL 8: [Wir haben] ja dann auch diese zusätzlichen didaktischen Materialien […], die wir dann halt auch im Bedarfsfall einsetzen können, wobei das halt im Moment noch wenig geschieht. Aber das habe ich ja auch schon gesagt, aus Zeitgründen einfach. (SL 8, 00:19:53-2)

SL 8: Und diese didaktischen Materialien, muss ich ehrlich sagen, ich habe sie bis jetzt nur überflogen. Machte mir einen ganz guten Eindruck, aber mehr kann ich dazu noch nicht sagen. (SL 8, 00:32:15-2)

Die Äußerung der Lehrkraft 12 belegt die vorangegangene Vermutung, dass zunächst ein neutrales Interesse an den didaktischen Materialien bei wenigen Lehrkräften vorhanden war. Die Inhalte sprachen bei der ersten oberflächlichen Rezeption jedoch nicht das Interesse an und somit erfolgte keine weitere Verwendung. Lediglich das Schulleitungsmitglied 8 erwähnte die Sinnhaftigkeit der didaktischen Materialien, um sich über die Anforderungen eines kompetenzorientierten Unterrichtens zu informieren. Die Nutzung geschähe jedoch nur im „Bedarfsfall". Dieser Bedarfsfall tritt vermutlich äußerst selten ein und ist stark von dem Interesse der Lehrkraft an den Inhalten abhängig, denn selbst der Schulleitungsvertreter 8, der grundsätzlich eine positive Einschätzung darlegte, hat das Material nur überflogen, so dass keine intensive Rezeption und Verwendung der Inhalte für den eigenen Unterricht erfolgte.

Zusammenfassend haben von 19 teilnehmenden Lehrkräften nur drei das didaktische Material gelesen, aber es wurde nie im Sinne einer Unterrichtsentwicklung genutzt. Demnach haben die Materialien bislang keine Bedeutung für die Lehrerschaft. Dies ist als bedauerlich einzuschätzen, da die didaktischen Materialien oftmals wertvolle und von den Lernstandserhebungen unabhängige Informationen enthalten, die eventuell für das gesamte Fachkollegium hilfreich sein könnten. Aus diesem Grund wären sie prinzipiell auch für die Schulleitungen von Interesse, um die Lehrkräfte in Hinblick auf kompetenzorientiertes Unterrichten zu motivieren und die Materialien an die Fachschaften weiterzugeben.

SL 6: Also ich denke, da kommt nicht so viel an. Es gibt ja hier zu diesen Erhebungen immer so eine CD, wo, glaube ich, Material drauf ist, wo auch, glaube ich, Vorschläge gemacht werden, was man machen kann. Das wird aber auch schon eher dezent wahrgenommen, würde ich mal sagen. (SL 6, 00:33:09-5)

SL 7: Also wir geben es natürlich an die Kollegen weiter, aber ich habe ehrlich gesagt bis jetzt noch nicht gehört, dass Kollegen das so unbedingt einsetzen. (SL 7, 00:35:30-6)

Folglich nahmen die Schulleitungen durchaus die Existenz solcher Angebote wahr und reichten sie teilweise auch an das Kollegium weiter. Jedoch konnte in den Interviews nicht festgestellt werden, dass Schulleitungsmitglieder, die an einer Lernstandserhebung selbst bislang nicht teilgenommen haben, sich ebenfalls mit den didaktischen Materialien be-

schäftigt haben. Vielmehr wurde lediglich registriert, dass die Materialien von den Kollegen äußerst geringfügig genutzt wurden.

9.1.2 Erfassung der Rezeptionsintensität mittels Kategorienbildung

Insgesamt konnten bei allen befragten Lehrkräften und Schulleitungen, die an den Lernstandserhebungen selbst teilgenommen haben, eine Form der Testrezeption festgestellt werden. Erwartungsgemäß unterschieden sich die Intensität dieser Nutzungsphase und die Rezeptionsaspekte. Um einen Überblick über die Verteilung der Rezeptionsintensität innerhalb der Befragten zu gewinnen, wurden vier Kategorien gebildet, denen das Rezeptionsverhalten jeder einzelnen Person zugeordnet wurde (vgl. Tabelle 9):

Kategorien zur Beschreibung der Rezeptionsintensität		
Kategorie	**Beschreibung**	**Zugeordnete Personen**
keine Rezeption	Es wurde keine Rezeption der Tests und der Rückmeldungen vorgenommen. Nach der Korrektur und Ergebniseingabe erfolgte keine weitere Beschäftigung.	-
geringe Rezeption	Die Rezeption war nur geringfügig, indem lediglich der durchschnittliche Gesamtwert der Klasse ODER die Testinhalte betrachtet wurden.	LK 1, LK 3, LK 5, LK 9, LK 11, SL 9
mittlere Rezeption	Es erfolgte eine Rezeption der nach Aufgaben und Kompetenzbereichen differenzierten Ergebnisse ODER eine Beschäftigung mit den Anforderungen einzelner Items beziehungsweise mit den didaktischen Materialien.	LK 8, SL 4, SL 5, SL 8, SL 12
intensive Rezeption	Es erfolgte eine Rezeption der nach Aufgaben und Kompetenzbereichen differenzierten Ergebnisse UND eine Beschäftigung mit den Anforderungen einzelner Items beziehungsweise mit den didaktischen Materialien.	LK 2, LK 4, LK 6, LK 7, LK 10, LK 12, SL 1, SL 3

Tabelle 9: Kategorien zur Erfassung der Rezeptionsintensität

Es wurde eine Einteilung in keine, geringe, mittlere und intensive Rezeption vorgenommen. Die Beschreibungen dieser Kategorien spiegeln lediglich die Spannbreite des anhand der Interviews ermittelten Rezeptionsverhaltens wieder und können nicht als allgemeingültig interpretiert werden. Es ist durchaus denkbar, dass bei einer anderen Probandenauswahl eine noch intensivere Rezeption hätte festgestellt werden können, so dass die Kategorienbildung daraufhin hätte angepasst werden müssen. Die Kategorien sollen lediglich das Verständnis über die beobachtete Rezeptionsintensität erleichtern sowie einen Überblick über die in der Untersuchung erhaltenen Informationen ermöglichen.

Aus der Tabelle 9 und den vorigen Ausführungen geht hervor, dass die Rezeption individuell höchst verschieden verläuft. Unter den reinen Lehrkräften rezipierten fast ebenso viele die Tests und die Rückmeldungen gering wie intensiv. Aufgrund einer relativ homogenen Verteilung auf die Kategorien 1 bis 3 lässt sich nur schwer eine Tendenz zu einer Intensitätsstufe ableiten. Lediglich unter den Schulleitungsvertretern lässt sich ein leicht positiver Trend erkennen. Als zufriedenstellend kann konnotiert werden, dass eine Rezeption stets erfolgte und bei einem Großteil der Befragten eine mittlere bis intensive Rezeption vorlag. Wie bereits dargelegt wurde, wird die Intensität der Nutzung von verschiedensten Faktoren, wie der zur Verfügung stehenden Zeit, der intrinsischen Motivation und dem Interesse, beeinflusst. Diese individuellen und äußeren Bedingungen sowie deren Auswirkungen auf den Nutzungsprozess werden in Abschnitt 10 vertieft vorgestellt.
Zuletzt soll an dieser Stelle angemerkt werden, dass die Rezeption zu verschiedenen Zeitpunkten stattfindet und nicht als ein zeitlich zusammenhängender Prozess begriffen werden kann. Einige Rezeptionsaspekte, wie das Durchführungs- und Korrekturmanual oder die Testinhalte, standen bereits während oder unmittelbar nach der Testdurchführung verstärkt im Interesse, während die Ergebnisse überwiegend erst nach Erhalt der Rückmeldungen rezipiert werden konnten.

9.2 Nutzungsphase der Reflexion

Dem Rezeptionsprozess schließt sich im Idealfall unmittelbar die Reflexion als weitere Phase in der Nutzung der Lernstandserhebungen an. Hierzu zählen eine Bewertung der Ergebnisse sowie der Tests, ein Vergleich mit den bisherigen Leistungen und das Ergründen der für die Ergebnisse verantwortlichen Ursachen. In dieser Nutzungsphase setzt ebenso die Kommunikation mit verschiedenen Personengruppen ein. Die Reflexion ist ein wesentliches Element für eine gewinnbringende Nutzung und stellte in den Interviews einen zentralen Gesprächsgegenstand dar.

9.2.1 Reflexionsaspekte

Bevor die Reflexion der Befragten hinsichtlich des Nutzens und der Intensität bewertet werden kann, ist eine Analyse der Reflexionsaspekte notwendig. Hierbei ist zum einen die Aufgabenreflexion zu nennen. Wie bereits in Abschnitt 9.1.1 dargelegt, wurden die Testinhalte von einigen wenigen Lehrpersonen rezipiert. In der Reflexion erfolgte eine Bewertung

der Aufgaben hinsichtlich ihrer Aufgabenformulierungen, der Itemschwierigkeit, den Durchführungs- und Korrekturbedingungen sowie der Lehrplanadäquatheit. Die Reflexion dieser Aspekte setzte ebenso wie die Rezeption bereits unmittelbar mit der Durchführung der Lernstandserhebung ein und ist daher der Ergebnisreflexion teilweise vorgelagert. Zum anderen wurden weitere Aspekte von den Probanden reflektiert, die im Folgenden dargelegt werden:

- Schülerverhalten,
- Zufriedenheit mit den Testergebnissen,
- Stärken und Defizite der Lerngruppe,
- Anregungen für die Weiterarbeit sowie
- Aussagekraft der Testergebnisse.

Schülerverhalten

Von zwei Lehrkräften wurde das Schülerverhalten reflektiert, wie an den folgenden Aussagen ersichtlich wird:

> *LK 3: Und was ich schon bemerkenswert fand und das war eine Rückmeldung für mich, das war, wie sie sich in einer Prüfung verhalten haben und das war sehr gut. Also das hätte ich gar nicht gedacht, dass diese-, also für mich waren es einfach kleine Leute, dass die da wirklich zwei Stunden konzentriert durchhalten. Und man muss sich ja durch diese vielen Bögen und Aufgaben und so, man muss sich ja schon irgendwie managen. Und das haben sie prima hingekriegt. (LK 3, 00:17:08-4)*
>
> *LK 7: Und ich habe, wenn ich es so höre, habe ich gedacht, ok, die waren relativ ruhig. Ich habe es geschafft, sie im Vorfeld zu beruhigen, was einem dann natürlich auch ein bisschen bestätigt. Weil ich auch gehört habe, dass andere Kinder ganz panisch waren. (LK 7, 00:19:59-8)*

Die Betrachtung des Schülerverhaltens erfolgte zunächst unabhängig von der Reflexion der späteren Ergebnisse. Jedoch bestand nach Ansicht der Lehrpersonen ein unmittelbarer Zusammenhang zwischen dem Schülerverhalten und der dargelegten Leistung. Aufgeregtheit habe sich beispielsweise auf die Konzentration der Schüler ausgewirkt. Ein immenser Umfang des Testmaterials habe eine Einschüchterung der Schüler bewirkt, was wiederum die Motivation und Leistungsfähigkeit beeinflusst habe.

Aus den obigen Aussagen lässt sich zudem eine Neugierde der Lehrkräfte auf die Reaktionen der Schüler in einer außergewöhnlichen Situation ableiten. Die Lernenden werden selten mit externen Tests konfrontiert, so dass die Beobachtung des Verhaltens während dieser Unterrichtssituation im Schulalltag nicht vorgenommen werden kann. Die Reflexion dieses Aspekts bezieht sich daher insbesondere auf die Bereiche der Selbst- und Arbeits-

kompetenz und kann neuartige Erkenntnisse für die Lehrkraft wie auch für die Schüler selbst liefern.

Zufriedenheit mit den Testergebnissen

Neben den bereits vorgestellten Betrachtungsgegenständen bezog sich die Reflexionsphase bei einem Großteil der teilnehmenden Lehrkräfte (zwölf von 19 Befragten) auf eine profunde Wertung der Schülerergebnisse. Dies drückte sich in allgemeiner Zufriedenheit beziehungsweise Unzufriedenheit mit den Leistungen der Schüler aus.

> *LK 6: Also ich muss ganz ehrlich sagen, ich habe mich sehr über das Ergebnis meiner Klasse gefreut im Verhältnis zum Referenzwert, die waren schon deutlich drüber. (LK 6, 00:07:20-0)*
>
> *SL 4: Also wenn wir jetzt diese Säulendiagramme sehen, können wir sagen, ok, es ist ja alles in Ordnung. Wie vorhin schon gesagt. Wir hätten es gerne ein bisschen besser, wissen aber, woran es liegt. Das ist ja klar. (SL 4, 00:32:27-7)*
>
> *SL 7: Insofern ist es natürlich erst mal für die Schule, für Lehrer und auch für Eltern beruhigend, wenn sie feststellen können - und das muss einfach für unsere Schule das Ziel sein oder das ist das Ziel -, dass wir auf jeden Fall im Landesdurchschnitt liegen oder besser. (SL 7, 00:11:06-0)*

Die Zufriedenheit mit den Schülerleistungen löste gewissermaßen eine Erleichterung bei den Lehrkräften aus, da die Schüler gut abgeschnitten haben. Dies ist eine bedeutsame Analyse, denn mit der Erfüllung der Erwartungen zu den Schülerleistungen ist ebenfalls eine Zufriedenheit mit der eigenen Arbeit als unterrichtende Person verknüpft. Die Erkenntnisse aus den Rückmeldungen repräsentieren daher nicht nur die erreichten Kompetenzen der Lernenden, sondern zugleich zu einem gewissen Grad die Leistung und Professionalität der Lehrperson.

Die Zufriedenheit mit der eigenen Arbeit ist nicht ausschließlich für die Selbstreflexion von Bedeutung. Vielmehr wird auf diese Weise die Lehrkompetenz auch nach außen gespiegelt. Innerhalb der Schule werden die einzelnen Klassenergebnisse oftmals in der Lehrerschaft miteinander verglichen und die Schulleitungen nehmen die Leistungsabweichungen zwischen einzelnen Lerngruppen ebenfalls wahr (vgl. Abschnitt 9.2.5). Daher impliziert die Zufriedenheit mit den Testergebnissen zugleich eine Erleichterung, sich als professionelle Lehrkraft gegenüber Kollegen, Vorgesetzten und den Eltern behauptet zu haben. Die Testergebnisse sind selbstverständlich nicht der alleinige Indikator für die Lehrkompetenz, doch wurden sie im Fall der Zufriedenheit als eine positive Bestätigung wahrgenommen.

Aus den obigen Interviewzitaten wird des Weiteren deutlich, wie diese Zufriedenheit mit den Testergebnissen entstanden ist. Überwiegend erfolgte dazu lediglich das Betrachten des Vergleichs der durchschnittlichen Klassenleistung mit dem korrigierten Landesmittel-

wert. Folglich diente der soziale Vergleich bei der Reflexion erneut als zentrale Bewertungsnorm. Die große Anzahl der Befragten, die auf diesen Vergleich bereits bei der Rezeption und nun auch bei der Reflexion Wert gelegt haben, bestätigen dies. Die zentrale Frage bei der Analyse der Rückmeldungen schien bei den Lehrkräften zu lauten: „Liege ich mit meiner Klasse über oder unter dem Landesmittelwert?“ Die Formulierung dieser Leitfrage wurde bewusst gewählt, da sie sowohl die Bedeutung des sozialen Vergleichs als auch die Verknüpfung mit dem eigenem Erfolg verdeutlicht.

Der reine Vergleich der Klassenergebnisse mit dem Landesmittelwert stellt jedoch eine sehr oberflächliche Analyse dar, die keine inhaltliche Bewertung der Leistung ermöglicht. Letztlich bleibt es offen, was ein „gutes“ Abschneiden tatsächlich bedeutet. Lediglich eine Vertreterin der Schulleitung nahm im Bereich der Zufriedenheit eine kriteriale Analyse vor:

> *SL 11: Aber was ich schon gut finde, […] immerhin die ganze Klasse sitzt […] in dem Kompetenzbereich als Punkt, als grauer. […] Und damit sind wir eigentlich ziemlich hoch, würde ich mal sagen. Sie haben ihnen eins, zwei, drei, vier, fünf, sechs, sieben, acht Pünktchen, sind wir auf dem B2-Niveau und C1. Da sind wir schon auf dem Oberstufenniveau.(LK 11 + SL 11, 00:11:39-4)*

Die zitierte Person analysierte folglich die Verteilung der Ergebnisse in den einzelnen Kompetenzstufen und konnte auf diese Weise eine kriteriale Bewertung der Ergebnisse vornehmen. Dass sich ein beachtlicher Teil der Lerngruppe in der achten Jahrgangsstufe annähernd auf dem Oberstufenniveau befand, nahm sie als eine bedeutsame Erkenntnis wahr, was zu einer persönlichen Zufriedenheit mit den Testleistungen führte.

Stärken und Defizite der Lerngruppe

Als einen weiteren Reflexionsaspekt bei neun befragten Personen ist das Erkennen von Stärken und Defiziten sowie von Anregungen für den eigenen Unterricht anzuführen.

> *LK 7: Ja, das habe ich mir so grob durchgelesen und versucht für mich raus zu filtern, da musst du etwas machen und das ist ok. (LK 7, 00:10:01-0)*
>
> *LK 9: Gut, tja man kann halt nochmal sehen, wo es bei den Schülern hängt, was für Aufgabentypen und eventuell auch welche Inhalte, ob es jetzt um, was weiß ich, Algebra geht oder um Geometrie oder irgendwie solche Dinge. Das kann man dann schon ablesen. (LK 9, 00:16:08-4)*
>
> *LK 10: Also sie gibt mir - und das ist das Wichtigste für mich oder für uns als Klasse -, diese Lernstandserhebung und ihre Ergebnisse geben uns wertvolle Hinweise auf das, was noch verbessert werden muss bei den Schülern. (LK 10, 00:09:33-9)*
>
> *SL 3: Und dann sieht man ganz deutlich bei den Kompetenzen, was man nicht gemacht hat. Also wenn man jemand ist, der nie Aufgaben in dem Metier gemacht hat, wird gerade im Argumentationsbereich untergehen. Und das finde ich für einen selbst spannend zu sehen, weil dann sieht man auch so ein bisschen, wo man in seinem Unterricht sich weiterentwickeln kann und ja, das finde*

ich den Hauptteil. […] Und da ist es auch interessant zu sehen, es gibt Themenbereiche, die in Schule oftmals runterfallen. Also in Mathematik aus Zeitmangel, aus welchen Gründen auch immer. […] Und das sieht man dann da auch sehr deutlich. (SL 3, 00:06:15-7)

Die Reflexion umfasste somit einerseits die Analyse, in welchen Bereichen Schwächen und Stärken zu erkennen sind, und andererseits die Fragestellung, wie diese Ergebnisse zu begründen sind. Die Analyse der Stärken einer Lerngruppe wurde wie der Reflexionsaspekt „Zufriedenheit" als eine persönliche Rückmeldung von den Lehrkräften zu ihrer geleisteten Arbeit und somit als eine positive Bestätigung wahrgenommen. Jedoch unterblieb stets eine Reflexion über die Hintergründe und die Ursachen der guten Leistung. Vielmehr wurden die Erkenntnisse lediglich angenommen und nicht hinterfragt. Ein Handlungsimpuls ist daraus bei keinem Befragten entstanden.

Bei defizitären Ergebnissen stellte es sich umgekehrt dar: Die Leistungen wurden nur geringfügig in Bezug zu der eigenen Arbeit als Lehrperson gesetzt. Dieser Aspekt wird in Abschnitt 9.2.3 bezüglich der Attribuierung von Ergebnissen nochmals aufgegriffen. Allerdings setzte bei erkennbaren Schwächen im Gegensatz zu den Stärken generell eine stärkere Ursachenanalyse zur Begründung der Resultate bei den Lehrkräften ein. Wie aus den Zitaten entnommen werden kann, entstanden durch die Analyse von Defiziten zudem Handlungsimpulse, um die Klasse in den Bereichen der Leistungsschwächen gezielt zu fördern.

Der soziale Vergleich mittels des Landesmittelwerts fungierte hierbei erneut als die zentrale Bezugsnorm. Er nahm eine deutlich größere Bedeutung als Bewertungsmaßstab der Leistungen ein als die Erkenntnisse aus der eigenen Korrektur. Somit bestätigt sich nochmals eine enorm starke Konzentration der Lehrkräfte auf den sozialen Vergleich. Für eine umfassende Analyse der Stärken und Schwächen der Lerngruppe wäre jedoch eine intensivere Aufgabenreflexion bezüglich der Zielsetzung und Anforderung der Items notwendig. Diese trat bei der Reflexion der Stärken und Defizite gänzlich in den Hintergrund.

Anregungen für die Weiterarbeit

Des Weiteren gewannen die Lehrkräfte während der Reflexion bereits Anregungen für die weitere Unterrichtsgestaltung.

LK 2: [Auch] sogenannte nicht-lineare Texte, Schemata, Statistiken und Ähnliches, und dass wir viel stärker als Lehrer darauf achten müssen, mit solchen Textformaten zu arbeiten und sie zu versprachlichen. Also von da aus ist die Rückmeldung für mich einerseits bestätigend, zum Teil aber auch aufschlussreich, weil es aufmerksam macht auf Defizite, die ich in meinem Unterricht so noch nicht in den Blick genommen habe. (LK 2, 00:07:12-5)

SL 1: Aber ich habe den Eindruck, dass die Gewöhnung der Schüler an das Sprachverstehen, dass die variabler gemacht werden. Bislang war es so, dass die Lehrer

zwar ihre eigene Sprache eingesetzt haben, das war so ein Bild für die Schüler. […] Und jetzt stellt man fest, bei den Lernstandserhebungen kommt auf einmal […] ein ganz anderer Sprachduktus rein. […] Das bedeutet, dass die Variabilität an dieser Stelle deutlich erhöht werden muss. (SL 1, 00:20:48-2)

Dementsprechend wurden vorrangig neue Inhalte und Unterrichtsgegenstände, wie Textformen und neuartige Medien, anhand derer Kompetenzen entwickelt werden können, als Anregungen wahrgenommen. Als ein weiteres Beispiel wurde von einer Lehrkraft der Kompetenzbereich „Hörverstehen" in Deutsch angeführt, den sie bislang noch nicht in den Unterricht integriert hatte. Interessanterweise sind diese Anregungen nicht ausschließlich mit der Reflexion der Rückmeldung gewonnen worden. Sie wurden vor allem bereits bei der Testdurchführung und bei der Auseinandersetzung mit den Testinhalten in Form einer Aufgabenreflexion während der Korrektur generiert. Den Auslöser für die Anregungen stellten somit die Testinhalte und -formate dar, welche die Lehrkraft als neuartig einstufte und daraufhin die Motivation entwickelte, sich intensiver damit zu beschäftigen. Demgegenüber wurden die Anregungen für die Weiterarbeit, die in den Rückmeldeberichten formuliert wurden, überhaupt nicht als einen Handlungsimpuls für den eigenen Unterricht von den Lehrpersonen interpretiert. Eine Begründung hierfür kann folgender Äußerung entnommen werden:

SL 4: *Und dass Binnendifferenzierung wichtig ist, das steht jetzt überall, das wusste ich vorher. […] Aber ich finde das, das ist so pauschal. […] Dann lese ich das, denke ich, ja gut. Wie geht es weiter? Was soll ich damit? (SL 4, 00:40:25-0)*

Folglich scheint der Hinweis „Binnendifferenzierung" keine ausreichende Wirkung zu erzielen, da er standardisiert formuliert ist. Das Schulleitungsmitglied 4 äußerte im weiteren Verlauf, dass eine Lehrkraft eher konkrete individuelle Hinweise für einzelne Schüler benötigen würde, um daraus Anregungen und Erkenntnisse für die weitere Förderung zu gewinnen. Diesbezüglich würde die Rückmeldung nicht genügend aufschlussreiche Informationen bieten, die auf die jeweilige Leistung der Lerngruppe oder einzelner Schüler angepasst sind.

Aussagekraft der Testergebnisse

Bei sechs befragten Personen konnte als eine weitere Reflexionstätigkeit das Hinterfragen der Ergebnisse hinsichtlich ihrer Aussagekraft konstatiert werden, welches sich auf verschiedene Aspekte der Lernstandserhebungen und ihrer Rückmeldungen bezog.

LK 11: *Im Mittelwert haben die hohen Schüler die Schwachen auf den Mittelwert gezogen. Die Klasse ist sehr unterschiedlich natürlich. (LK 11 + SL 11, 00:07:52-0)*

LK 8: *Ich kann sagen: "Du hast 75 Prozent erreicht." […] War halt so, aber was für Konsequenzen ziehe ich daraus? Ziehe ich jetzt die Konsequenz: "Ach ich habe ja 75 Prozent, ich muss mich gar nicht anstrengen!" Oder andere: "Ach, ich hab nur 75 Prozent." (LK 4, 00:14:30-5)*

SL 4: Also das sind ja Bausteine, die irgendwie zusammengesetzt werden. Das macht ja keiner persönlich. […] Wobei ich das dann bei diesem Writing am lächerlichsten fand, weil wir da alle geschwommen sind in der Bewertung dieser Texte, die die Kinder geschrieben haben. Und dann kriege ich eine objektiv aussehende Rückmeldung, wo die Kinder sich befinden. Ich halte das für Augenwischerei. […] Ich denke, das ist so eine Scheinobjektivität, die da aufgebaut wird. (SL 4, 00:40:25-0)

Eine kritische Betrachtung erfuhr die statistische Methode der Ergebnisberechnung in Form von Durchschnittswerten. Einerseits kann die Lehrkraft auf diese Weise keine Informationen über einzelne Schüler erhalten, da der Durchschnittswert die Klasse zu einem homogenen Wert degradiert. Dementsprechend lässt er keine Aussagen über die Leistungsheterogenität und die Spannweite der Leistungen innerhalb der Klasse zu, was nicht der Realität entspricht. Demgegenüber muss erwähnt werden, dass in der Rückmeldung ebenfalls ein Streuungsdiagramm enthalten ist. Wie bereits dargestellt wurde (vgl. Abschnitt 9.1.1), stand dieses Element nicht im Vordergrund der Rezeption, denn die Lehrkräfte fokussierten sich stark auf den sozialen Vergleich. Doch die Bewertung eines Durchschnittswertes im Verhältnis zum Landesmittelwert erschien den Lehrkräften unklar, da durch die Kompensation von positiven und defizitären Leistungen ein Ergebnis vermittelt wird, welches so nicht existiert. Vermindert wird diese geringe Aussagekraft zusätzlich, wenn die Differenz zwischen dem Klassenergebnis und dem Landesmittelwert nicht signifikant deutlich ist.

Des Weiteren wurde von der Lehrkraft 4 generell der soziale Vergleich hinterfragt. Es blieb dem Befragten unklar, wie eine Leistung über oder unter dem Landesmittelwert zu interpretieren ist. Für eine spezifische Interpretation würden zusätzliche Informationen, wie zur Itemkonzeption oder eine Berücksichtigung des Lernprozesses, benötigt. Sicherlich ist es positiv zu deuten, wenn sich die Lerngruppe oberhalb des durchschnittlich erreichten Wertes in Hessen befindet. Doch ist diese Leistung als ausreichend oder als besonders gut zu werten? Die Lehrkraft 4 verweigerte sich der pauschalen Bewertung der Schülerleistung und stellte damit erneut die Notwendigkeit einer kriterialen Deutung heraus.

Einen zusätzlichen Aspekt stellt die Angabe der Ergebnisse in Form von prozentualen Werten dar. Da die Lernstandserhebungen nicht wie herkömmliche Klassenarbeiten konzipiert sind, sondern wissenschaftlichen Gütekriterien genügen müssen, kann der übliche Bewertungsmaßstab nicht angewandt werden, nach dem eine Leistung von 75 Prozent beispielsweise als „befriedigend“ zu beurteilen wäre. Folglich fehlt den Lehrpersonen der zugehörige Bewertungsmaßstab, um die Ergebnisse angemessen und für sich verständlich interpretieren zu können. Es können daher keine Schlüsse zu den Kompetenzständen der Schüler aus den prozentualen Angaben abgleitet werden. Hierfür wäre eine genaue Verortung der

Leistungen auf den Kompetenzstufenmodellen erforderlich, die ein inhaltliches Bewerten der Leistungen ermöglichen würde.

Als letzten Kritikpunkt in Bezug auf die Aussagekraft der Ergebnisse wurde vom Schulleitungsmitglied 4 die fragwürdige Objektivität der Tests angeführt. Da in diesem Fall vor allem die Korrektur der Lernstandserhebungen als nicht objektiv bewertet wurde (vgl. Abschnitt 11.3), wurde die generelle Aussagekraft der Rückmeldungen infrage gestellt. Infolge der fehlenden Objektivität sei es nach Ansicht des Schulleitungsvertreters fraglich, inwiefern die tatsächliche Leistung eines Schülers überhaupt erfasst wird.

9.2.2 Vergleiche der Ergebnisse mit weiteren Leistungsdaten der Schüler

Ein wesentliches Element der Reflexion ist der Vergleich mit den bisherigen Bewertungen der Schüler im zurückliegenden Unterricht. Die Lehrkraft setzt dabei die Testwerte in Bezug zu den eigenen Leistungseinschätzungen, welche aus langfristigen Benotungen und Beobachtungen, wie beispielsweise prozessbezogenen Beurteilungen oder Klassenarbeiten, resultieren. Grundsätzlich können hierbei zwei Ergebnisse vorliegen: Entweder nimmt die Lehrperson eine Bestätigung ihrer bisherigen Einschätzungen durch die Testergebnisse wahr oder es treten Differenzen zwischen dem im Test dargelegten Kompetenzerwerbsstand und den aus dem Unterricht resultierenden Bewertungen auf. Für den ersten Fall ist die Betrachtung der folgenden Interviewzitate von Bedeutung:

> *LK 3: Eigentlich habe ich mich schon bestätigt gefühlt in meiner Einschätzung. Jetzt Kinder, die Probleme haben, Texte richtig zu verstehen, beim Lesen aufzufassen, worum es da geht, die haben dasselbe Bild da in dem Test auch gezeigt. (LK 3, 00:16:14-4)*
>
> *LK 10: Also ich habe es insofern in Verbindung gebracht, dass unmittelbar davor eine Klassenarbeit geschrieben wurde und man also sehen konnte, dass oft dieselben Schüler ähnliche Probleme hatten. Das hat man so ein bisschen wiedererkannt. (LK 10, 00:08:09-7)*
>
> *SL 5: Was für mich interessant war, das hat eigentlich so ein bisschen das Leistungsvermögen der Klasse, das ich sonst bei schriftlichen Arbeiten habe, wiedergespiegelt. Also der, der am schlechtesten abgeschnitten hat, ist auch so im Deutsch eigentlich so der schlechteste Schüler. Und der Beste im Test ist auch mein bester Schüler im Unterricht. (SL 5, 00:05:14-6)*

Insgesamt äußerten neun der 19 befragten Personen, die an den Lernstandserhebungen teilgenommen hatten, dass ihre bisherigen Einschätzungen durch die Testergebnisse zumindest grob bestätigt wurden. Diese Übereinstimmung zeigte sich beispielsweise in einem Vergleich mit konkreten Notengebungen, wie sie von der Lehrkraft 10 und dem Schulleitungsmitglied 5 vorgenommen wurde. Bei einem Großteil der Probanden beruhte der Ver-

gleich der Testdaten mit der persönlichen Einschätzung jedoch nicht auf konkreten Noten, sondern auf der allgemeinen Wahrnehmung des Leistungsprofils der Klasse. Dementsprechend konnten einige Lehrpersonen bereits im Vorfeld erfolgreich prognostizieren, welche Schüler das leistungsstarke Viertel bei den Lernstandserhebungen bilden und welche Schüler schwächer abschneiden würden. Die Form des Vergleichs mittels persönlicher Wahrnehmungen ist generell unpräziser und erfolgt weniger intensiv als der Vergleich mit konkreten Daten. Jedoch schien bei der Mehrheit der Befragten kein Interesse an einer tiefergehenden Vergleichsanalyse zu bestehen.

Dass die Ergebnisse in etwa als äquivalent zum sonstigen Leistungsprofil zu charakterisieren sind, bestätigt zum einen die Lehrkraft in ihrer professionellen Wahrnehmung. Wenn die eigenen Beurteilungen mit denen eines externen Tests übereinstimmen, scheint die Lehrperson folglich die Schüler in ihrem Können adäquat einschätzen zu können. Zum anderen unterstreicht es zugleich die Güte des Tests an sich, da dieser augenscheinlich eine valide Erfassung der Schülerleistungen ermöglicht.

Die Gründe für die Bestätigung der Einschätzung konnten selten benannt werden. Lediglich der längerfristige Einsatz als Lehrperson in den getesteten Klassen wurde als Ursache angeführt, da ein intensives Kennenlernen einzelner Leistungsprofile bereits stattgefunden hätte und ein gefestigtes Erwartungsbild bei den Lehrkräften vorhanden sei. Demgegenüber bewiesen einige Lehrkräfte, wie sogleich dargelegt werden wird, dass trotz klarer Bewertungseinschätzungen der Vergleich grundlegend neue Erkenntnisse mit sich bringen kann. Folglich ist die Intensität, mit welcher die Vergleiche vorgenommen werden, ein den Erkenntnisgewinn stark beeinträchtigender Faktor.

Des Weiteren muss angemerkt werden, dass eine unbewusste Beeinflussung der Voreinschätzungen auf die Korrektur der Lernstandserhebungen nicht ausgeschlossen werden kann. Dies steht im Zusammenhang mit einem Maß an Professionalität der Lehrkraft, neutral an die Bewertung eines solchen Tests heranzutreten. Es ist allerdings durchaus möglich, dass Leistungserwartungen auf die Korrektur einwirken und dementsprechend die Ergebnisse zu einem gewissen Teil steuern. Diese Vermutung kann in Bezug auf eine Lehrkraft entkräftet werden, bei der ein Teil der Korrekturen von einer externen Praktikantin übernommen worden waren. Deren Ergebnisse spiegelten ebenfalls das Leistungsbild wieder, so dass von einer objektiven Korrektur ausgegangen werden kann. Dennoch kann die Möglichkeit einer unbewussten Einflussnahme auf die Testergebnisse aufgrund eigener Leistungseinschätzungen nicht gänzlich verworfen werden.

Neben einer Bestätigung können bei diesen Vergleichen ebenso Differenzen zwischen den Testdaten und der persönlichen Wahrnehmung für überraschende Erkenntnisse sorgen.

LK 2: Es ist überraschend, dass es gerade-, jetzt sage ich mal, die fleißigen, braven Mädchen, die meistens aufpassen, die mit dabei sind, dass die gerade bei diesen engen Vorgaben von Richtig oder Falsch versagt haben. (LK 2, 00:09:37-6)

LK 2: Zum anderen macht es natürlich auch aufmerksam, dass es Schüler gibt […] - das betrifft häufig Jungs, weil sie disziplinarisch eben auffällig sind, weil sie faul sind, weil sie unter ihren Möglichkeiten bleiben -, dass die dann an einem solchen Test plötzlich zeigen, was in ihnen steckt. Also da werde ich schon aufmerksam, dass manche überraschend gut und manche überraschend schlecht abschneiden. (LK 2, 00:11:13-7)

LK 3: Aber was mir schon aufgefallen ist, dass auch plötzlich Gute und sehr Gute unter ihrem Niveau geblieben sind, einfach weil sie überfordert waren von der Menge und von der Zeit, in der die sich zu konzentrieren hatten. (LK 3, 00:09:20-9)

LK 5: Ich denke jetzt gerade speziell an eine Schülerin, wo es wirklich umwerfend war, wie viel sie da zustande gebracht hat und auf welchem Niveau. Also das fand ich dann doch schon nochmal bestätigend und auch nochmal so positiv, sehr positiv überraschend. (LK 5, 00:06:19-5)

Die Abweichungen von der eigenen Einschätzung betrafen größtenteils einzelne Schüler und weniger die gesamte Klasse. Bei diesen Vergleichen bildeten nicht konkrete Benotungen die Grundlage, sondern vielmehr Verhaltensaspekte der Schüler im Unterricht. Dies bestätigt den massiven Einfluss der Verhaltenswahrnehmung auf die Leistungswahrnehmung. In dessen Konsequenz bieten die Lernstandserhebungen die Chance, neue Perspektiven zum Leistungsvermögen einzelner Schüler zu erhalten.

Als Ursachen für die Abweichungen zwischen den Testwerten und den eigenen Erwartungen wurden vorrangig drei Begründungen angeführt: Zum einen seien die Schüler an andere Aufgabenstellungen gewöhnt und folglich sei ihnen der Umgang mit den Itemformaten und Anforderungen erschwert gewesen, was schwächere Resultate als angenommen zur Folge gehabt habe. Zum anderen wurden die Rahmenbedingungen der Testdurchführung und -auswertung angeführt, indem die Lernstandserhebungen beispielsweise zu umfangreich gewesen seien oder die fehlende Wertung der Rechtschreibung bei der Korrektur andere Ergebnisse befördert habe als im herkömmlichen Unterricht. Drittens wurde in Form einer kritischen Selbstreflexion erörtert, dass gewisse Kompetenzen im Unterricht wenig gefördert würden, so dass ein mögliches Potenzial bislang noch nicht sichtbar geworden sei. Generell teilten insgesamt zehn befragte Lehrpersonen von den eigenen Einschätzungen abweichende Ergebnisse mit. Davon haben jedoch lediglich fünf Lehrkräfte die Ursachen für diese Differenzen analysiert.

Aus den Interviews ging zudem hervor, dass bei acht Befragten weiterführende Vergleiche vorgenommen wurden. Beispielsweise erfolgten Gegenüberstellungen zwischen den Testergebnissen und den Bewertungen der Parallelarbeiten, welche im gleichen Fach geschrieben wurden. In einem Fall wurden zudem die Testdaten zwischen den einzelnen Fächern

einer Jahrgangsstufe miteinander in Beziehung gesetzt, um daraus ein Gesamtbild für das Abschneiden der Schule zu erhalten. Öfter erfolgte hingegen der Vergleich zwischen Parallelklassen. Dies hat den weiteren positiven Effekt, dass zugleich die Kommunikation der Kollegen über die Lernstandserhebungen generell gefördert wird. In einer Schule wurde dies zugleich für eine Art informeller Evaluation einer Lerntalentklasse genutzt, indem die Differenz zwischen den Ergebnissen dieser Klasse und den anderen Lerngruppen betrachtet wurde. Trotz dieser Vergleiche konnten die Befragten hierbei keine neuen Erkenntnisse gewinnen. Es ist jedoch auffällig, dass der Vergleich verschiedener Leistungserhebungen miteinander ausschließlich von Schulleitungsvertretern beziehungsweise von Lehrkräften, welche die Funktion eines Fachsprechers ausüben, wahrgenommen wurde. Die Betrachtung eines größeren Kontextes, bei der Analysen über die eigene Lerngruppe hinaus vorgenommen werden, schien somit nicht im Interesse der übrigen Lehrkräfte zu stehen.

Insgesamt konstatierten sieben Lehrende, mittels der Rückmeldungen keine entscheidend neuen Erkenntnisse über das Leistungsvermögen der Klasse erhalten zu haben:

> *LK 3: Also ich hatte hinterher das Gefühl, ich habe jetzt nichts Neues über meine Schüler dazugelernt, was ich nicht schon vorher gewusst hätte. […] Also ja, es gab wirklich keine gravierend neuen Aufschlüsse. (LK 3, 00:16:14-4)*

> *SL 4: Also wenn man sieht, ok, das ist so im Mittel, dann kann man sagen: "Gut, was soll es?" Dann steht irgendwo, man müsste ein bisschen binnendifferenziert-. Das weiß man ja vorher! Also es sind ja keine neuen Erkenntnisse, die man da gewinnt. (SL 4, 00:13:42-6)*

Alle weiteren Äußerungen diesbezüglich wiesen den gleichen Inhalt auf. Die Gruppe der Befragten, welche keinen Erkenntnisgewinn feststellen konnten, deckt sich mit denjenigen Personen, deren Voreinschätzungen bestätigt wurden. Folglich kann ein Zusammenhang festgestellt werden: Wenn die eigenen Erwartungen an das Leistungsbild der Lerngruppe durch die Ergebnisse in den Lernstandserhebungen bestätigt werden, erhält die Lehrkraft keine für sie neuartigen Informationen. Zu untersuchen bleibt der Einfluss dieses Zusammenhangs auf die Handlungsmotivation in der Phase der Aktion sowie auf die Bewertung des Nutzens der Lernstandserhebungen (vgl. Abschnitte 9.3 und 11.6).

9.2.3 Attribuierung der Ursachen

Zu einer erkenntnisreichen Reflexion zählt neben der Analyse der Testergebnisse auch die intensive Ergründung der zugehörigen Ursachen. Nur auf diese Weise können die Unterrichtsprozesse mit den Testdaten in einem Wirkungszusammenhang betrachtet werden. Wie in Abschnitt 5.5 dargelegt wurde, wird zwischen internaler und externaler Attri-

buierung unterschieden. Dementsprechend werden im Folgenden die Aussagen der Befragten ebenfalls diesen beiden Kategorien zugeordnet.

9.2.3.1 Internale Attribuierung

Bei der internalen Attribuierung werden die Ergebnisse mit der Qualität der eigenen Arbeit als Lehrperson verknüpft, so dass die Verantwortung für die Testleistung selbst übernommen wird. Die folgenden Interviewzitate ermöglichen einen Einblick, in welchen Bereichen eine internale Attribuierung erfolgte:

LK 2: Und das überraschend Bessere, glaube ich, liegt daran, dass ich im Unterricht bei Besprechung von Texten generell großen Wert auf das innere Erleben, also auf die Innenperspektive lege und wir das immer wieder besprochen haben. Und das hat sich zum Teil positiv ausgewirkt. (LK 2, 00:07:12-5)

LK 2: Ich glaube, […] dass sie eher bislang bei mir gelernt haben, wenn ich eine Auffassung über das Verhalten einer Figur oder einen Sachverhalt aus dem Text begründe, dann muss ich nicht ein Wort oder eine Zeile oder nur einen Satz angeben, sondern dann gibt es einen Kontext. […] Und das ist etwas, was ich mir gewissermaßen so als Geschichte erkläre. (LK 2, 00:11:13-7)

LK 6: Also ich habe einiges kennengelernt und habe das dann in [Unterrichtsbesuchen] und auch sonst im Unterricht dann eben umgesetzt und habe auch versucht, auch mit einer vollen Stelle das überwiegend so fortzuführen. Und ich denke, darin liegt dann auch der Erfolg der Klasse, besonders beim Leseverstehen auf den verschiedenen Kompetenzstufen. (LK 6, 00:15:42-7)

SL 4: Also dieser Nachfolger [des Schulbuchs] […] bereitet die Schüler eigentlich unheimlich gut auf Hörverstehen vor. Da hatten die auch ganz gut abgeschnitten. (SL 4, 00:08:50-8)

Aus den Interviewausschnitten geht hervor, dass bei positiven Ergebnissen internale Attribuierungen vorgenommen wurden. Dies wurde beispielsweise damit begründet, dass im Unterricht bereits auf die gleichen Aspekte und Aufgabenanforderungen Wert gelegt wurde, wie sie später getestet wurden. Dementsprechend waren die Schüler ohne ein direktes Üben für die Lernstandserhebungen gut vorbereitet und erzielten entsprechende Ergebnisse. Dies lässt den Schluss zu, dass diejenigen Lehrkräfte bereits kompetenzorientierte Zielsetzungen im Unterricht berücksichtigt hatten beziehungsweise das verwendete Lehrbuch dahingehend eine Orientierung bot. Bei Betrachtung der beruflichen Laufbahn derjenigen Befragten, die solche Äußerungen trafen, konnte festgestellt werden, dass drei der vier Probanden entweder kürzlich das Referendariat abgeschlossen hatten oder als Ausbilder in einem Studienseminar tätig waren. Folglich waren sie mit der Kompetenzorientierung theoretisch wie praktisch vertraut, was sich teilweise in den Ergebnissen ihrer Schüler in positiver Weise wiederspiegelte.

Generell wurde insbesondere von den Schulleitungsvertretern oft die Auffassung vertreten, dass die Lehrkraft vorrangig für die Ergebnisse der Schüler verantwortlich sei und in zweiter Linie erst die durchschnittliche Leistungsfähigkeit der Lerngruppe. Beispielsweise orientiere sich ein effektiver Unterricht nach Äußerung eines Schulleitungsvertreters an komplexen Herausforderungen für die Schüler, welche zurückliegende Themen einbinden und somit eine Vernetzung verschiedener Aspekte des Faches ermöglichen würden. Dies entspricht zugleich dem kompetenzorientierten Konzeptionsansatz der Lernstandserhebungen, indem nicht kurzfristig erworbenes Wissen, sondern langfristig aufgebaute Kompetenzen überprüft werden. Wird der Unterricht diesem Anspruch gerecht, sind dementsprechend auch gute Leistungen im Test zu erwarten. Die Verantwortung der Lehrperson für die Testresultate ist auf diese Weise in einem direkten Zusammenhang mit ihrer Professionalität zu deuten. Letztlich führt eine internale Attribuierung bei positiv zu bewertenden Ergebnissen zu der Bestätigung, den Unterricht den Anforderungen gemäß zu gestalten und mit dieser Vorgehensweise erfolgreich zu sein.

Aber auch defizitäre Ergebnisse wurden teilweise internal begründet, wie beispielweise aus dem angeführten Zitat von Lehrkraft 2 hervorgeht. Demzufolge kann durchaus eine Diskrepanz zwischen den eigenen pädagogischen Vorstellungen und den Unterrichtsanforderungen zu den Lernstandserhebungen entstehen, indem wie im exemplarischen Fall auf Aspekte im Unterricht besonderen Wert gelegt wurde, die im Test wiederum für die Schüler hinderlich waren, da kürzere Antworten erwartet wurden. Folglich schnitten die Schüler negativer ab. Die Lehrkraft 2 beurteilte im weiteren Verlauf des Gesprächs ihr Vorgehen im Unterricht als das pädagogisch sinnvollere und kritisierte in diesem Kontext die Testkonzeption. Der Unterschied zwischen Lern- und Testaufgaben (vgl. Abschnitt 4.3.1.2) ist hierbei von Bedeutung, welcher der Lehrperson nicht bewusst war. Vielmehr nahm sie den Test ebenfalls als eine Lernsituation wahr. Aus der Aussage der Lehrkraft 2 lässt sich zudem schließen, dass die Schüler bei den Lernstandserhebungen neuartigen Anforderungen gegenüberstehen, die den Prozessen des Fachunterrichts teilweise entgegenstehen.

Als weitere Begründungen für defizitäre Ergebnisse wurden die Vernachlässigung der getesteten Themen im Unterricht (was erneut zur Verantwortung der Lehrkraft zu zählen ist) sowie die kurze Einsatzdauer als Lehrperson in der jeweiligen Lerngruppe genannt. Bei letzterem Argument wurde von einer reduzierten Verantwortlichkeit der Lehrenden ausgegangen, da die Weichen bereits in vorangegangen Lernprozessen gelegt worden seien. Des Weiteren konnten nicht alle Testleistungen eindeutig internal attribuiert werden, wie folgende Aussagen aufzeigen:

SL 8: Ich meine, so ein schlechtes Abschneiden bei der Lernstandserhebung sagt ja nicht unbedingt […] etwas darüber aus, dass man schlechten Unterricht gemacht hat oder so. Dann sind vielleicht wirklich die Gegebenheiten in der Klasse so, dass es halt nicht anders ging. (SL 8, 00:29:52-2)

SL 12: [Ich] habe festgestellt, dass gewisse Dinge, die die Schüler schon vor geraumer Zeit bearbeitet haben, eigentlich recht ordentlich so bearbeitet hatten. Und dass das, was ich ganz frisch gemacht habe, nämlich diese Geradengleichungen, Steigungen herausfinden und dann das Ganze spiegeln und dann daraus Schlüsse ziehen, das ist ja ganz frisch gewesen bei mir. Und da dachte ich, och, kein Problem. Das packen die alle und da gibt es hundert Prozent. Pustekuchen, war gar nicht der Fall! Da waren plötzlich wirklich Fehler aufgetreten, die ich mir definitiv im Moment nicht erklären konnte. […] Ich bin ratlos, warum das nicht gelungen ist! (LK 12 + SL 12, 00:20:37-6)

Am Beispiel des Schulleitungsmitglieds 12 wird deutlich, dass zwar eine internale Begründung angestrebt wurde, sich jedoch keine sinnvollen Ursachen für die Testergebnisse erklären ließen. Es wäre natürlich möglich, dass die Thematik beispielsweise noch nicht genügend gefestigt war oder das Item andere Zielsetzungen aufwies als die im Unterricht eingesetzten Aufgaben. Doch eine eindeutige Benennung war nicht möglich. Demnach griff eine ausschließlich internale Attribuierung zu kurz und befriedigte bei diesen spezifischen Teilergebnissen das Erkenntnisinteresse nicht. Dies wurde auch vom Schulleitungsvertreter 8 ausgedrückt, indem aus schlechten Ergebnissen nicht automatisch die Schlussfolgerung eines schlechten Unterrichts gezogen werden könne. Vielmehr seien weitere Aspekte von Bedeutung, die der externalen Attribuierung zuzuordnen sind und ebenso als Einflussfaktoren auf die Ergebnisse wirken würden.

Resümierend soll angemerkt werden, dass eine internale Attribuierung nur bei sieben an den Tests teilnehmenden Lehrkräften festzustellen war. Jedoch bezog sich diese internale Attribuierung nicht ausschließlich auf positive Ergebnisse, sondern wurde durchaus auch bei Unzufriedenheit mit den Testleistungen vorgenommen.

9.2.3.2 Externale Attribuierung

Im Rahmen der externalen Attribuierung bezogen sich die Äußerungen der Befragten auf die folgenden Bereiche: Ursachen auf Schüler- beziehungsweise Klassenebene, Ursachen auf Schulebene sowie Ursachen auf Testebene.

Schüler- bzw. Klassenebene

Auf die Schüler- und Klassenebene bezogen wurden zwei verschiedene Argumentationen für die Testergebnisse angeführt. Die folgenden Zitate dienen dazu, zunächst einen dieser Ursachenkomplexe exemplarisch zu veranschaulichen:

> *LK 4: Die Kinder der Lerntalentklasse zeichnen sich im Allgemeinen durch eine höhere Selbstständigkeit in der Organisation ihrer Lernprozesse aus. Dies sollte sich in der durchschnittlichen Leistung der Klasse zeigen, da es eine größere starke Gruppe als in herkömmlichen Klassen gibt und schwächere Schüler sich eher im 3 - 4er-Notenbereich aufhalten. (LK 4, Nachgespräch per E-Mail)*
>
> *LK 9: Wir haben zum Beispiel ein Phänomen, dass unsere Bläserklasse, die also in Klasse 5 und 6 gemeinsam so eine Art Blasorchester halt aufbauen, Musikinstrumente lernen, das ist halt die Klasse, die in Mathe jetzt deutlich oben war. […] Das hängt sicherlich auch damit zusammen, dass halt vielleicht die Elternhäuser natürlich ausgewählt sind, weil es eben auch etwas kostet und eben zusätzlich ist. (LK 9, 00:23:53-2)*
>
> *SL 4: Das ist ja klar. Wenn wir in Klassen Schüler haben, die zum Beispiel in Deutsch nicht so fit sind, weil sie aus einem Elternhaus kommen, wie aus Kambodscha. (SL 4, 00:32:27-7)*
>
> *LK 5: Also wir haben drei Schüler zum Beispiel in der Gruppe, die haben [eine Leserechtschreibschwäche], die bekommen eigentlich immer mehr Zeit. Das war ja nicht eingeplant hier bei solchen Sachen. Und das hat sich schon bemerkbar gemacht, gerade bei den produktiven Aufgaben. Die hätten länger benötigt, um da schreiben zu können und deshalb waren da einige Sachen frei. Ich denke, die hätten die Aufgaben bewältigt mit zehn, fünfzehn Minuten mehr Zeit, wie sie es auch sonst im Unterricht bekommen. (LK 5, 00:05:11-4)*

Demzufolge wurde die Klassenzusammensetzung von sechs befragten Personen als ein bedeutender Einflussfaktor angeführt. Jede Lerngruppe ist aufgrund ihrer Heterogenität in ihrem Leistungsniveau voneinander verschieden, was sich nicht ausschließlich bei den Lernstandserhebungen, sondern auch in Parallelarbeiten oder anderen externen Tests, wie dem Mathematikwettbewerb, wiederspiegelt. Als Spezialfall kann zum einen eine Lerntalentklasse in der Schule 4 angeführt werden, die bereits in Abschnitt 9.2.2 erwähnt wurde. Diese zeichne sich in ihrer pädagogischen Konzeption insbesondere durch eine verstärkte Selbstständigkeit der Schüler in der Organisation ihrer Lernprozesse aus. Im Rahmen der Lernstandserhebungen wurde deutlich, dass nach Aussage der Lehrkraft 4 das leistungsstärkste Viertel der Schüler bei der Quartilierungsauswertung bessere Punktzahlen aufwies als in den Parallelklassen. Diese Erkenntnis wurde mit der Besonderheit der Lerntalentklasse begründet. Interessanterweise beurteilte die zugehörige Schulleitungsvertreterin dies anders, indem keine signifikanten Unterschiede zwischen den einzelnen Lerngruppen festzustellen seien. Allerdings ist die Schulleitung 4 sehr kritisch gegenüber den Lernstandserhebungen eingestellt und derjenigen Gruppe der Befragten zuzuordnen, welche die Aussagekraft der Ergebnisse stark hinterfragt haben. Zugleich ist offen, wie intensiv die Schullei-

tung 4 und die Lehrkraft 4 diesen Vergleich der Ergebnisse zwischen den Parallelklassen vorgenommen haben. Da die Lehrperson 4 als Klassenlehrer der Lerntalentklasse eingesetzt war, kann davon ausgegangen werden, dass sie über eine differenzierte Einschätzung zum Leistungsvermögen seiner Klasse verfügte.

Ähnliche Beobachtungen stellte die Lehrkraft 9 im Kontext einer Bläserklasse fest, wobei sie nicht primär die Unterrichtprozesse als Begründung heranzog, sondern die Besonderheit der Lerngruppe hinsichtlich ihres familiären sozialen Hintergrunds. Auch das Schulleitungsmitglied 4 führte an, dass der sozioökonomische Status der Familien einen massiven Einfluss auf das Leistungsvermögen und auf die Testergebnisse einzelner Schüler ausüben würde. Dieser Aspekt wurde von dem Befragten als eine spezifische Benachteiligung einzelner Schüler bei den Lernstandserhebungen gedeutet, da sich die Testinhalte in den Sprachen stark auf grammatikalische Kompetenzen sowie auf den Sprachwortschatz konzentrieren würden. Diese Begründung kann teilweise mit der Berücksichtigung des sozioökonomischen Status im Kontext des fairen Vergleichs anhand der sogenannten Bücherfrage entkräftet werden. Hierbei müssen die Schüler durch Ankreuzen angeben, wie viele Bücher sich in ihrem Elternhaus befinden, beispielweise 20-100 Bücher oder 100-200 Bücher. Inwiefern diese Fragestellung es erlaubt, realistische Annahmen über den tatsächlichen sozioökonomischen Status abzuleiten, darf kritisch hinterfragt werden, denn es stellt lediglich ein Indiz für das akademische Profil einer Familie dar. Zudem ist unklar, inwiefern die Schüler die Bücherfrage tatsächlich wahrheitsgemäß beantworten. Dennoch wird eine Benachteiligung einzelner Schüler durch den fairen Vergleich in einem gewissen Maß relativiert. Der faire Vergleich kommt jedoch ausschließlich bei einer Gegenüberstellung der Ergebnisse zum korrigierten Landesmittelwert zum Tragen. Bei der Betrachtung des Sofortberichts und der tabellarischen Auswertung kann kein Bezug auf die Hintergrundinformationen zu den Schülern vorgenommen werden. Folglich hatten diese Ergebnisdarstellungen eine verminderte Aussagekraft in der Wahrnehmung der Lehrperson, so dass die zur Verfügung gestellten Informationen an Relevanz verloren.

Aus diesem Grund müssen die Lernbedingungen der Schülerschaft bei der Reflexion berücksichtigt werden. Zudem sollte die Schülerperspektive Beachtung finden. Da die Tests nicht auf die Bedürfnisse und den individuellen Lernstand einer Klasse angepasst sind, erfahren einige Schüler einen schulischen Misserfolg. Demgegenüber kann aber angemerkt werden, dass die Lehrkraft anhand der Testergebnisse spezifische Hinweise für die individuelle Förderung dieser Schüler erhalten kann. Dieser Blickwinkel wurde jedoch von den Befragten nicht geteilt.

Als zweiter Ursachenkomplex im Rahmen der externalen Attribuierung auf Schüler- und Klassenebene wurden die Motivation und die Konzentration der Schüler während des Tests angeführt:

> *LK 6: [Da] gab es Probleme die einzelnen Schlössertypen zuzuordnen. Da dachte ich dann, ok, vielleicht waren sie dann beeinflusst von Aufgabe 1, die sie extrem leicht fanden, wo sie dann dachten: "Ok, wir brauchen uns doch nicht so zu konzentrieren, dann machen wir Aufgabe 2 auch locker." Ich denke, das war so ein Problem. (LK 6, 00:17:28-3)*
>
> *LK 9: Es sind ein paar, die das halt ein bisschen auf die leichte Schulter nehmen, weil es halt nicht zählt, und dann halt relativ schlecht abschneiden, also nicht so wie sonst. (LK 9, 00:06:24-1)*
>
> *LK 11: Da gibt es ein paar, die hatten keine Lust. (LK 11 + SL 11, 00:08:17-2)*

Demnach waren die Schüler aufgrund des permanenten Wechsels der Itemschwierigkeiten teilweise leichtfertig an schwerere Aufgaben herangetreten, haben deren Anforderungen unterschätzt und dementsprechend schlechter abgeschnitten. Andere Lehrkräfte führten die fehlende Motivation bei den Schülern als Begründung an. Ursache hierfür sei die Nicht-Bewertung der Lernstandserhebung als eine Note, so dass die Schüler keine maximale Leistungsbereitschaft aufgewiesen hätten. Dies verzerre die Ergebnisse der Lernstandserhebungen, da unklar bleibe, über welchen Kompetenzstand die Schüler tatsächlich verfügen und inwiefern dieser von der Testleistung abweicht. Als Bestätigung für diese Begründung wurde der Vergleich mit den bisherigen Noten der Schüler und der persönlichen Einschätzung der Lehrkraft angeführt, indem von der Klasse bessere Ergebnisse erwartet worden seien. Des Weiteren stellten die Befragten bei einigen Schülern eine Nervosität fest, da die Bearbeitung eines hessenweiten externen Vergleichstests Druck ausgeübt hätte und zu viel in einzelne Aufgaben hineininterpretiert worden sei.

Die externalen Attribuierungen bezüglich der Konzentration und Motivation der Schüler wurden ausschließlich bei defizitären Ergebnissen vorgenommen. Zugleich wurde stets eine Wirkungskette aufgebaut: Die Testkonzeption oder die Nicht-Bewertung führte zu einer demotivierenden Reaktion bei den Schülern, was wiederum schlechtere Ergebnisse zur Folge hatte. Folglich wurden die primären Ursachen auf der Testebene angesiedelt, währenddessen Desinteresse oder mangelnde Konzentration lediglich als Resultate hiervon interpretiert wurden.

Schulebene

Auf Schulebene fanden nach Auswertung der durchgeführten Interviews nur äußerst selten Attribuierungen statt. Einmal wurde die Schulform des Gymnasiums als Ursache für die Leistungen angeführt. Ein Gymnasium stelle hohe Leistungsanforderungen, die zu erfüllen

seien und dementsprechend seien die Unterrichtsprozesse größtenteils leistungsorientiert zu bewerten. Demzufolge sei es obligatorisch, als Gymnasium zufriedenstellend bei den Lernstandserhebungen abzuschneiden.

Die zweite Äußerung in diesem Bereich beschrieb ein konkretes Problem in der Schulorganisation. Aufgrund eines längeren krankheitsbedingten Ausfalls einer Lehrkraft erhielt die getestete Klasse lediglich Vertretungsunterricht, der von mehreren Lehrpersonen durchgeführt wurde. Dies hätte Auswirkungen auf die Qualität und die Nachhaltigkeit der Lernprozesse gehabt, was sich wiederum in schlechteren Testergebnissen im Vergleich zu den Parallelklassen widergespiegelt hätte.

Testkonzeption

Neben Ursachenbegründungen auf Schüler-, Klassen- und Schulebene wurden externale Attribuierungen bezüglich der Testkonzeption vorgenommen. Zum einen wurden die inhaltliche Gestaltung und die Zielsetzung der Lernstandserhebungen von einem Großteil der Befragten (zehn von 19 an den Tests teilnehmenden Lehrpersonen) als maßgebliche Einflussfaktoren auf die Bearbeitung der Items und somit auf die Schülerergebnisse angeführt, wie folgende Interviewausschnitte exemplarisch veranschaulichen:

> *SL 4: Die werden mit Aufgabenformaten konfrontiert, die sie vorher nicht gesehen haben […] und dann tauchen die plötzlich auf. Ja dann wundere ich mich nicht, wenn die Kinder das nicht können, wenn ich das vorher, wenn die das erste Mal das sehen und vorher ganz andere Dinge gewöhnt sind. (SL 4, 00:51:58-3)*
>
> *SL 1: Es gibt bestimmte Themen, die haben die Schüler zu dem Zeitpunkt noch nicht durchgenommen. Dann können sie diese Aufgaben nicht und es ist klar, dann ist es Zufall, ob sie richtig oder falsch gelöst haben. Oder das liegt zumindest nicht ganz in unserer Hand, ob sie es richtig, ob sie es konnten oder hätten gekonnt haben können. (SL 1, 00:14:12-5)*
>
> *LK 7: Sie mussten sich jetzt erst mal in dieses Indian English rein hören. […] Also das, fand ich, war eine sehr schwere Aufgabe. […] Und da haben sie dann auch am schlechtesten abgeschlossen im Vergleich zu den einzelnen Listening-Aufgaben. (LK 7, 00:06:51-3)*

Der Inhalt einzelner Items ist dahingehend ein zentraler Begründungsgegenstand für die Ergebnisse, inwiefern der Test den Themen und Anforderungsniveaus der jeweiligen Jahrgangsstufe entsprechend konzipiert ist. Teilweise sind diese Attribuierungen durchaus berechtigt. Wenn beispielsweise Thematiken getestet werden, die bislang nicht in den Lehrplänen enthalten waren, sind die Schüler voraussichtlich nicht in der Lage, diese Aufgabe zu lösen. Diese Items werden aber dennoch mit null Punkten versehen, so dass es sich negativ auf die Gesamtbewertung auswirkt. Es ist jedoch durchaus legitim, dass die Lernstandserhebungen weiterführende Aspekte überprüfen, die nur ein geringer Teil der Lerngruppe

lösen kann. Zu einer Testauswertung, die wissenschaftlichen Gütekriterien genügen soll, gehören ebenso Items, die Leistungsstärken genügend differenziert erfassen können. Dieses testtheoretische Argument ist aber für die individuelle Lehrkraft unerheblich. Vielmehr hatten die defizitären Ergebnisse bei solchen Items eine grundsätzlich negativere Bewertung der Testgestaltung zur Folge, so dass die Verantwortung für die Teilergebnisse von sich gewiesen wurde.

Alternativ wäre in solchen Fällen eine Reflexion darüber denkbar, welche Zielsetzungen sich hinter der Aufnahme eines solchen Items im Test verbergen. Darauf aufbauend könnte eine Betrachtung stattfinden, wie sie eventuell in die eigene Unterrichtsentwicklung integriert werden kann. Diese Form einer positiv ausgerichteten Reflexion ist bei keinem Befragten erfolgt.

Gleiche Reaktionen konnten bei Aufgaben beobachtet werden, deren Inhalte erst später im Unterricht behandelt werden sollten. Aufgrund der Terminierung der Lernstandserhebungen im Februar beziehungsweise März eines Schuljahres ist es möglich, dass in einigen Schulen bestimmte thematische Einheiten bereits Unterrichtsgegenstand gewesen sind und in anderen aufgrund eines Schulcurriculums oder wegen Absprachen innerhalb des Jahrgangsteam erst später im Unterricht aufgegriffen werden würden. Die Lehrkräfte beurteilten dies als problematisch, da die Schüler die Items nicht lösen konnten, dies aber wiederum bewertet wurde und somit eine Verfälschung des Ergebnisses zu erwarten war. Die Lehrkraft müsste in einem solchen Fall die betreffenden Items bei der Reflexion der Ergebnisse außer Acht lassen. Dennoch wirkte sich die Bepunktung einer nicht erfolgten Bearbeitung auf die Ermittlung des korrigierten Landesmittelwerts und somit auf den sozialen Vergleich aus.

Eine Abschwächung der Reaktion, den Testinhalt als Ursache für die Ergebnisse zu deuten, lässt sich im Verhalten der Lehrkraft 7 erkennen. Diese beurteilte ein Item der Lernstandserhebung zwar ebenfalls als schwer und sah darin ein schlechteres Abschneiden ihrer Schüler begründet. Allerdings wertete sie dies neutral und verband damit keine Abwertung der Testqualität. Vielmehr empfand sie als selbstverständlich, dass ein solcher Test neben durchschnittlichen Aufgabenschwierigkeiten ebenso sehr leichte oder sehr schwere Aufgaben beinhaltet.

Als eine weitere Begründungslinie ist die Unbekanntheit bestimmter Aufgabenformate anzuführen, mit denen die Schüler in den bisherigen Lernprozessen noch nicht konfrontiert worden waren. Da dies eine weitere neue Anforderung an die Lernenden darstellte, wirkte es sich nach Aussage der Befragten ebenfalls negativ auf die Ergebnisse aus.

Neben den Attribuierungen inhaltlicher Testaspekte wurden die Rahmenbedingungen der Lernstandserhebungen, wie die Durchführungsregularien, als Ursachen für die Resultate der Schüler von den Lehrkräften betrachtet:

> *LK 3: Aber was mir schon aufgefallen ist, dass auch plötzlich Gute und sehr Gute unter ihrem Niveau geblieben sind, einfach weil sie überfordert waren von der Menge und von der Zeit, in der die sich zu konzentrieren hatten. Und für die Kinder hätte die Hälfte an Aufgaben dicke genügt. (LK 3, 00:09:20-9)*
>
> *LK 5: Und die haben nicht einfach aufgegeben, sondern sie hatten keine Zeit mehr. Und das ist so ein Zeichen für mich, sie haben es noch nicht mal probiert. Sie konnten es gar nicht probieren, weil die Zeit gar nicht mehr da war. (LK 5, 00:10:55-2)*
>
> *LK 4: Die Hörverstehenstexte […] waren sehr schnell vorgetragen. Und die Schüler hatten auch wenig Zeit, über das, was sie da produzieren, großartig nachzudenken. Also es waren wirklich dann Dinge, wo man spontane Einfälle von Schülern erwartet, man aber doch weiß, dass es auch Schüler gibt, die eine sehr gute Leistung oder die Anforderung nach gewisser Zeit des Nachdenkens erreichen können. Dadurch hat man den Effekt, dass manche Schüler sehr schnell durch die Tests durchgegangen sind und da aber oberflächlich gearbeitet haben. Manche sehr intensiv gearbeitet haben, dafür aber den Test nicht ganz geschafft haben. Und dadurch verfälscht das in gewisser Weise natürlich auch die Rückmeldung. (LK 4, 00:16:01-3)*

Als eine zentrale Ursache für Fehler bei der Bearbeitung wurden von insgesamt fünf befragten Lehrkräften der hohe Umfang an Testitems sowie die Schnelligkeit, in welcher die Aufgabenstellungen zu lösen waren, betrachtet. Dies führte nach Aussage der jeweiligen Lehrkräfte zu einem Konzentrations- und Motivationsverlust bei den Schülern. Ein solches oberflächliches Arbeiten steht jedoch dem Qualitätsanspruch von Lernprozessen entgegen und hatte eine erhöhte Fehlerquote zur Folge. Falls dennoch Schüler intensiv an einzelnen Fragestellungen gearbeitet haben, reichte ihnen die zur Verfügung stehende Zeit zur Lösung des gesamten Tests nicht, was wiederum defizitäre Ergebnisse bewirkte. Die Schüler waren demnach sowohl vom Umfang der Tests, als auch von der zeitlichen Dauer überfordert, da die sonstigen schulischen Leistungsüberprüfungen in den Klassenstufen 6 und 8 jeweils maximal eine Unterrichtsstunde umfassen statt neunzig Minuten. Letztlich würden diese Aspekte die Ergebnisse verfälschen, denn neben dem Anwenden von erworbenen Kompetenzen werde eine Schnelligkeit in der Arbeitsweise gefordert, welche die notwendigen Denkprozesse maßgeblich erschwere. Es blieb für die Lehrkräfte unklar, inwiefern die Ergebnisse der Lernstandserhebungen das Leistungsvermögen der Lerngruppe tatsächlich wiedergespiegelten oder ob der Schüler die Aufgabe in einer anderen Situation richtig gelöst hätte.

Als ein weiterer Aspekt wurde die Gestaltung der Testhefte kritisch hinterfragt, da deren letzte Seite von vielen Schülern nicht beachtet worden sei und somit mehrere Items nicht

gelöst wurden. Dies wirkte sich massiv auf das Gesamtergebnis aus. Allerdings ist während der Durchführung der Lernstandserhebung die anwesende Lehrperson für die Einweisung der Schüler zuständig, zu der ebenfalls die Erläuterung des Testheftes zählt.

Im Kontext der externalen Attribuierung auf Testebene können zuletzt die Korrekturbestimmungen angeführt werden, die von den Lehrpersonen auch als Ursachen für Testergebnisse betrachtet wurden.

> *SL 5: Einige haben besser abgeschnitten, weil ja die Rechtschreibung nicht bewertet wird. (SL 5, 00:05:14-6)*
>
> *LK 11: Aber gut, je nachdem, wie ich es korrigiert habe. Ich habe es streng korrigiert, kommt nicht so viel bei raus. Ich habe mich erkundigt parallel zur anderen Achten und da wurde es nicht ganz so streng korrigiert. Da sitzen die dann natürlich ein bisschen höher. Was hat das für eine Aussagekraft? (LK 11 + SL 11, 00:05:01-4)*
>
> *LK 4: Was mich da gestört hat, […] ab einer gewissen Zahl fehlender Schüler ist es mir nicht mehr möglich, diese auch als fehlend anzugeben, sondern ich muss dann Schüler mit null Punkten einsetzen. Wenn ich die mit hundert Prozent der Punkte einsetze, egal mit welcher Zahl ich sie einsetze, es verzerrt mein Ergebnis am Ende. (LK 4, 00:35:12-5)*

Von dem Schulleitungsmitglied 5 wurde der Einfluss der fehlenden Bewertung der Rechtschreibung angesprochen. Da die Rechtschreibung insbesondere im Sprachunterricht während der Lernprozesse und in Leistungsüberprüfungen eine zentrale Rolle einnimmt, führte diese Regelung bei einzelnen Schülern zu massiven Abweichungen zwischen den bisherigen Beurteilungen und dem Ergebnis in der Lernstandserhebung. In den meisten Fällen war dies für den Lernenden als positiv zu werten, da sich ihre Punktzahlen aufgrund von Rechtschreibfehlern entgegen der herkömmlichen Praxis in Klassenarbeiten nicht reduzierten. Dies kann einen neuen Blickwinkel auf die Leistungsfähigkeit des Schülers ermöglichen. Die Lehrkräfte beurteilten es jedoch anders. Sie empfanden diese Korrekturbestimmung als unpädagogisch und den Zielen des Fachunterrichts entgegenstehend, was Auswirkungen auf die Wahrnehmung der Relevanz der Ergebnisse für die eigene Diagnose hatte. Letztlich war die Rechtschreibung kein Testgegenstand, doch inwiefern ist diese Vorgehensweise sinnvoll, wenn die Orthographie im Unterricht selbst einen bedeutsamen Raum einnimmt?

Als weitere Ursache wurde die Variabilität innerhalb der Korrekturanweisungen angeführt. Die Lehrperson könne je nach Auslegung der Musterlösung die Ergebnisse beschönigen oder im negativen Sinn beeinflussen. Es läge keine Eindeutigkeit in der Bewertungspraxis vor und dies führe zu einer Abschwächung des Gütekriteriums der Objektivität, indem unterschiedliche Lehrpersonen unterschiedliche Korrekturen vornähmen. Die Lehrkräfte, die diesen Aspekt anführten, bezogen zugleich eine kritische Position zu den Vorgaben durch

die Testkonstrukteure. Eine genauere Analyse der Bewertung der Tests durch die Lehrkräfte, zu welcher auch die Korrekturbestimmungen zählen, findet in Abschnitt 11 statt.
Letztlich wurden technische Rahmenbedingungen als Einflussfaktoren angeführt. Indem nur eine gewisse Anzahl an Schülern als fehlend markiert werden konnten, musste für alle weiteren abwesenden Schüler eine beliebige Punktzahl eingetragen werden. Dies verfälschte selbstverständlich in hohem Maße die Gesamtergebnisse der Klasse, die auf Durchschnittswerten beruhen. In einem solchen Fall kann keine realistische Punktzahl eingesetzt werden, da unklar ist, wie der entsprechende Schüler die Items gelöst hätte.
Resümierend kann festgestellt werden, dass externale Attribuierungen auf Testebene fast ausschließlich bei defizitären Ergebnissen vorgenommen und hierbei vielfältige Begründungen angebracht wurden. Eine Ausnahme bildete die fehlende Wertung der Rechtschreibung, die positivere Ergebnisse bewirkt hatte. In der Perspektive der Lehrkräfte wurde dies jedoch, wie dargelegt wurde, ebenfalls kritisch analysiert. Die Ursachenfindung auf Testebene ermöglichte es dem Lehrer, die eigene Verantwortung für die Schülerergebnisse (fast) vollständig abzugeben. Teilweise sind die Argumente durchaus berechtigt und einige Einflussfaktoren auf die Ergebnisse können nicht abgestritten werden. Andererseits wurden in einigen Fällen die Rahmenbedingungen dahingehend interpretiert, dass eine naheliegende internale Attribuierung vermieden wurde. Eventuelle Chancen für die eigene Unterrichtsentwicklung, die mit der intensiven Reflexion von negativen Ergebnissen verbunden sind, wurden durchgängig nicht erkannt.

9.2.4 Erfassung der Reflexionsintensität mittels Kategorienbildung

Zurückblickend betrachtet nahmen fast alle befragten Lehrkräfte und an dem Test selbst teilnehmende Schulleitungsvertreter eine Reflexion vor. Lediglich bei einer Person war diese Nutzungsphase nicht zu konstatieren. Die Gründe hierfür werden später analysiert. Ebenso wie bei der Rezeption wies der Prozess der Reflexion bei den Befragten unterschiedliche Dimensionen hinsichtlich seiner Nutzungsintensität auf. Um diese Dimensionen erfassen zu können, wurden äquivalent zur Rezeption erneut vier Kategorien entwickelt, denen das Reflexionsverhalten der einzelnen Lehrkräfte und an dem Test teilnehmenden Schulleitungsvertretern zugeordnet wurde (vgl. Tabelle 10).

Kategorien zur Erfassung der Reflexionsintensität		
Kategorie	**Beschreibung**	**Zugeordnete Personen**
keine Reflexion	Es wurde keine Reflexion in Form einer Analyse der Testergebnisse und des Begründens der Ergebnisse vorgenommen.	SL 9
geringe Reflexion	Die Reflexion war nur geringfügig, indem lediglich die Testergebnisse mit weiteren Leistungsdaten und -einschätzungen vergleichend analysiert wurden.	LK 9, LK 11, LK 12, SL 3, SL 5, SL 8
mittlere Reflexion	Im Rahmen der Reflexion erfolgte eine vergleichende Analyse der Testergebnisse mit weiteren Leistungsdaten und -einschätzungen UND eine Attribuierung zur Begründung der Ergebnisse.	LK 1, LK 3, LK 5, LK 8, SL 4, SL 12
intensive Reflexion	Im Rahmen der Reflexion erfolgte eine vergleichende Analyse der Testergebnisse mit weiteren Leistungsdaten und -einschätzungen UND eine Attribuierung zur Begründung der Ergebnisse. Des Weiteren wurden Erkenntnisse für die Unterrichtsentwicklung abgeleitet.	LK 2, LK 4, LK 6, LK 7, LK 10, SL 1.

Tabelle 10: Kategorien zur Erfassung der Reflexionsintensität

Es erfolgte eine Einteilung in keine, geringe, mittlere und intensive Reflexion. Die Analyse der Testergebnisse im Vergleich mit bisherigen Einschätzungen und Leistungsbenotungen stellte die Grundlage für eine weiterführende Reflexion dar und wurde daher als die niedrigste Nutzungsstufe angenommen. Darauf aufbauend erhöhte sich die Reflexionsintensität, indem die jeweiligen Ursachen und Begründungszusammenhänge abgeleitet wurden. Die höchste Reflexionsstufe zeichnet sich zusätzlich durch das Ableiten von Anregungen und Erkenntnissen zur Weiterentwicklung des Unterrichts auch. Die Kategorie „intensive Reflexion“ bildet daher die ideale Voraussetzung für den weiterführenden Nutzungsprozess im Bereich der Aktion. Erst wenn Entwicklungsmöglichkeiten erkannt werden, können diese auch umgesetzt werden.

Die Beschreibungen dieser Kategorien spiegeln lediglich die Spannbreite des anhand der Interviews ermittelten Reflexionsverhaltens wieder und können nicht als allgemeingültig interpretiert werden. Auch hier ist eine intensivere Reflexion theoretisch möglich, so dass die Kategorienbildung daraufhin entsprechend angepasst werden müsste. Die Kategorien sollen lediglich das Verständnis über die Reflexionsintensität erleichtern.

Aus der Tabelle 10 und den im Vorfeld gemachten Ausführungen geht hervor, dass die Reflexion individuell unterschiedlich intensiv verläuft. Es lässt sich in etwa die Gleichwertigkeit der Anzahl der zugeordneten Personen in den Bereichen der geringen bis intensiven Reflexion feststellen. Im Vergleich zu der Verteilung der Rezeptionsintensität ist erkennbar, dass die Schulleitungsvertreter zunehmend bei einer geringeren bis mittleren Nutzung zu verorten sind. Folglich scheinen die Lehrkräfte eher nach Ursachenerklärungen zu suchen und

Anregungen zu den Tests abzuleiten. Dies steht in einem Zusammenhang mit den Ergebnissen aus der Rezeptionsphase: Indem vor allem bei den Schulleitungsvertretern eine Positionierung der Klasse anhand des sozialen Vergleichs im Zentrum der Betrachtungen stand, wurde bei einer positiven Verortung der Lerngruppe oberhalb des Landesmittelwerts Zufriedenheit mit den Resultaten verbunden, woraus kein weiteres Erkenntnisinteresse entstand. Folglich beschränkte sich die Reflexion auf die Deutung des sozialen Vergleichs und eine Attribuierung der Ergebnisse erfolgte nur bei defizitären Ergebnissen.

Die Reflexion der vorgestellten Aspekte erfolgte innerhalb des Probandenstammes erwartungsgemäß unterschiedlich intensiv. Folgende Äußerungen lassen Rückschlüsse auf mögliche Ursachen für eine geringere Reflexion zu:

LK 4: Allerdings muss ich sagen, bedeutet die Auswertung des Tests auch viel Zeit, die ich so im Alltag, im Unterrichtsalltag mir nicht nehmen konnte. (LK 4, 00:05:14-1)

LK 6: Also, ich denke, wenn man jetzt ein halbes Jahr in der Klasse eingesetzt ist, dann ist es schwierig, daraus Rückschlüsse zu ziehen. (LK 6, 00:21:48-7)

SL 6: Also so eine systematische Beschäftigung dann mit den Hintergründen dieser Ergebnisse, die ist noch in den Kinderschuhen bei uns. (SL 6, 00:38:15-0)

Als einen wesentlichen Hindernisgrund wurde wie bei der Rezeption die nicht ausreichende zur Verfügung stehende Zeit angegeben. In der Tat erfordert eine intensive Auswertung der Rückmeldungen eine gewisse Zeitdauer, welche die Lehrkraft zunächst einmal aufzubringen bereit sein muss. Folglich sind erneut die Motivation und die Bedeutung, die den Ergebnissen zugemessen wird, ausschlaggebende Faktoren für die Nutzung. Eine solche intensive Auswertung ist notwendig, um gehaltvolle Erkenntnisse und Handlungsimpulse zu erhalten und die Lernstandserhebung für die Weiterentwicklung des Unterrichts zu verwenden.

Die Äußerung der Lehrkraft 6 offenbart zudem eine persönliche Distanz zwischen der Bewertung des eigenen Unterrichts und den vorliegenden Testergebnissen aufgrund des bislang kurzen Einsatzes in der Klasse. Allerdings sollte hierbei berücksichtigt werden, dass die Lernstandserhebungen im Februar beziehungsweise März eines Schuljahres durchgeführt werden. Die Lehrperson unterrichtete die Klasse somit im Regelfall bereits mindestens ein halbes Jahr und sollte die Leistungen der Schüler differenziert beurteilen können. Andernfalls kann die Reflexion der Testwerte dennoch bedeutsam sein, da sie Diagnoseinformationen bereitstellt, die zur Einschätzung der Lernenden dienlich sind.

Ein weiterer Hindernisgrund für eine Reflexion ist der Stellenwert, den die Lernstandserhebungen innerhalb einer Schule einnehmen. Wenn dieser relativ gering ist, wirkt sich dies erschwerend auf die Motivation für eine intensive Auswertung bei der einzelnen Lehrkraft aus. Hierbei ist insbesondere das Vorhandensein eines kollegialen Kommunikationskon-

zepts von Bedeutung. Bei einer ausführlichen gemeinsamen Besprechung der Ergebnisse würde der Reflexionsprozess enorm befördert werden.

Des Weiteren konnte festgestellt werden, dass insbesondere bei Lehrkräften, bei denen eine negative Bewertung der Tests hinsichtlich des Inhalts und der Rahmenbedingungen ersichtlich war, die Reflexion ebenfalls eingeschränkt verlief.

Des Weiteren wird aus der Zuordnung der Befragten zu den Kategorien bezüglich der Rezeption und Reflexion deutlich, dass eine geringe Rezeption keineswegs den Abbuch oder Verringerung der Intensität des Nutzungsprozesses zur Folge hat. Vielmehr waren die einzelnen Nutzungsphasen oftmals nicht konstant hinsichtlich ihrer Intensität bei den jeweiligen Befragten. Lediglich bei fünf Probanden konnte eine abnehmende Nutzung festgestellt werden. Davon betrug die Differenz in drei Fällen eine Intensitätsstufe. Bei den anderen beiden Befragten wurden eine intensive Rezeption, jedoch nur eine geringe Reflexion konnotiert, was eine Abschwächung um zwei Intensitätsstufen bedeutet. Als interessant stellte sich hierbei heraus, dass von diesen fünf Personen insgesamt vier Mitglieder der Schulleitung waren, was die oben angeführten Erkenntnisse zu der geringeren Reflexion bei Schulleitungsvertretern bestärkt. Als wesentlichen Grund für die weniger intensive Reflexion wurde der unzureichende Erkenntnisgewinn aus den Informationen der Rückmeldungen angegeben. Die Ursache hierfür ist wiederum die Fokussierung auf den sozialen Vergleich. Lediglich die Verortung der Lerngruppe war von Interesse, was eine einseitige Bewertung der Ergebnisse in den Kategorien „über oder unter dem Landesdurchschnitt“ begünstigte. Weiterführende Analysen waren für das jeweilige Informationsbedürfnis nicht von Belang. Indem die Voreinschätzungen auf diese Weise in allen Fällen bestätigt wurden, bewerteten die Befragten den Erkenntnisgewinn durch die Tests als sehr gering. Dies ist in der oberflächlichen Analyse des Probanden selbst begründet, wurde allerdings in dieser Weise nicht von ihm bewusst wahrgenommen. Durch den geringen Erkenntnisgewinn reduzierte sich die Bewertung der Tests als ein für sich persönlich nützliches Instrument und die Motivation zu einer weiterführenden Reflexion mit den Testinhalten oder -ergebnissen sank. Somit erschwerte die Konzentration auf den sozialen Vergleich massiv das Interesse und die Bereitschaft zu einer intensiveren Auseinandersetzung mit den Leistungen der Schüler.

Innerhalb dieser Gruppe, welche durch eine reduzierte Nutzungsintensität gekennzeichnet ist, ist der Schulleitungsvertreter 9 zu verorten, der als einziger Befragter keine Reflexion durchgeführt hat. Seine Äußerungen lassen sich sehr deutlich auf die Beschränkung auf den sozialen Vergleich und die oben beschriebene Wirkungskette beziehen. Er stellte heraus, dass lediglich die Verortung für ihn relevant sei und andere Informationen ihn nicht interessieren würden. Ein Vergleich mit Voreinschätzungen wurde ebenfalls nicht vorgenommen.

Des Weiteren bewertete er die Lernstandserhebungen und insbesondere die zugehörigen Materialien in ihrer Qualität als negativ, so dass die Erfahrungen aus der Rezeptionsphase sich erschwerend auf die Reflexion auswirkten und einen vorzeitigen Abbruch des Nutzungsprozesses zur Folge hatten. Inwiefern dies die Nutzungsphase der Aktion beeinflusste, wird in Abschnitt 9.3 beschrieben.

Bei insgesamt drei Befragten ließ sich demgegenüber eine zunehmende Nutzungsintensität von einer geringen Rezeption zu einer mittleren Reflexion feststellen. Bei diesen Probanden war eine fast ausschließlich externale Attribuierung auf Testebene zu beobachten, die vorrangig bei negativen Ergebnissen vorgenommen wurde. Eventuell ist dies die Folge einer geringfügigen Rezeption, indem aufgrund einer mangelnden Auseinandersetzung mit den Ergebnissen und Materialien die Verantwortung für Testleistungen der Schüler primär dem Test selbst zugeschrieben wurden.

Bei den restlichen elf Befragten lag eine gleichbleibende Nutzungsintensität in den Phasen der Rezeption und Reflexion vor. Auffällig ist hierbei die hohe Anzahl von fünf Befragten, die der intensiven Nutzungskategorie zuzuordnen sind. Demnach beförderte eine gründliche Rezeption eine intensive Reflexion in besonderem Maße. Indem die Testinhalte und Ergebnisse tiefgründig betrachtet wurden, entwickelten die Lehrkräfte Informationsbedürfnisse, um sich beispielsweise die Testleistungen erklären zu können. Dies führte zu einer erhöhten Motivation und Arbeitsbereitschaft im Bereich der Reflexion.

9.2.5 Kommunikation als Merkmal der Reflexionsphase

Die gewonnenen Erkenntnisse aus den Rückmeldungen der Lernstandserhebungen berühren zunächst den eigenen Unterricht einer Lehrkraft. Gemäß dem Zyklenmodell von Helmke (vgl. Helmke A. , 2004, S. 100) soll die Nutzung der Testergebnisse über diese Ebene hinausgehen, indem eine Kommunikation mit verschiedenen Personengruppen stattfindet. Dies erweitert nicht nur den Erkenntnisgewinn, sondern fördert zugleich die Reflexionsfähigkeit der Lehrperson und stößt mithilfe des Aufbaus und der Festigung von Kommunikationsstrukturen die Organisationsentwicklung innerhalb der Schule an. Dabei sind folgende Kommunikationspartner möglich:

- Kommunikation zwischen den teilnehmenden Lehrkräften, innerhalb einer Fachschaft beziehungsweise im Rahmen einer Gesamtkonferenz,
- Kommunikation zwischen der teilnehmenden Lehrkraft und der Schulleitung sowie innerhalb der Schulleitung,

- Kommunikation zwischen der teilnehmenden Lehrkraft und ihren Schülern,
- Kommunikation zwischen der teilnehmenden Lehrkraft und den Eltern,
- Kommunikation zwischen schulischen Akteuren und außerschulischen Personen beziehungsweise Institutionen.

In der qualitativen Untersuchung wurden Gesprächsanlässe und Intensität der Kommunikation über die Lernstandserhebungen nach diesen Bereichen differenziert ausgewertet. Die Ergebnisse werden im Folgenden dargelegt.

Kommunikation zwischen teilnehmenden Lehrkräften, innerhalb einer Fachschaft beziehungsweise im Rahmen einer Gesamtkonferenz

Die Kommunikation über die Lernstandsergebnisse und der Schülerresultate kann zum einen informell in Form zufälliger Gespräche zwischen einzelnen beteiligten Lehrkräften und zum anderen offiziell im Rahmen von Fachschaftssitzungen und Gesamtkonferenzen erfolgen. Für den inoffiziellen Austausch sind keine Kommunikationsstrukturen institutionalisiert. Aus diesem Grund wird diese Kommunikationsebene in der Regel ohne äußeren Druck wahrgenommen. Demgegenüber ist eine intrinsische Motivation der einzelnen Lehrkraft von umso größerer Bedeutung. Es muss ein gegenseitiges Kommunikationsbedürfnis vorhanden sein, damit ein Austausch entstehen kann.

Bei der Auswertung der Interviewaussagen konnte das Vorhandensein von Kommunikationsbedürfnissen bei der Mehrheit der Probanden festgestellt werden. Die folgenden Interviewpassagen konkretisieren diese Interessen inhaltlich:

> *LK 2: Es brauchte eigentlich idealerweise einer Art gemeinsamer Vorbesprechung aller Kollegen. (LK 2, 00:21:45-6)*
>
> *LK 2: Es wäre vielleicht auch insofern besser, wenn die, die das letzte Jahr gemacht haben, den Kollegien der nachfolgenden Klasse 6 oder 8 sagen: „Wir haben die Erfahrung gemacht." (LK 2, 00:29:22-8)*
>
> *SL 2: [Es] müsste eine institutionalisierte Rückmeldung erfolgen an beteiligte wie nicht-beteiligte Kollegen. (SL 2, 00:12:40-5)*
>
> *LK 10: [Da] sollte man nach der Auswertung sich definitiv zusammensetzen und schauen, welche typischen Fehler gemacht wurden. Auch schauen, ob zum Beispiel in einer anderen Klasse andere Fehler gemacht wurden in einer gewissen Häufigkeit, um dieses Fehlerpotenzial an dieser Stelle auch zu berücksichtigen für den eigenen Unterricht. (LK 10, 00:19:17-9)*

Aus diesen Aussagen wird ersichtlich, dass der Wunsch nach einem Austausch innerhalb des Kollegiums verschiedene Elemente des Nutzungsprozesses anspricht. Einerseits wurde die Sinnhaftigkeit einer gemeinsamen Absprache bezüglich der organisatorischen Durchführung von einem Probanden (Lehrkraft 2) erwähnt. Andererseits wurde der Nutzen einer

gemeinsamen Auswertung der Testergebnisse angeführt, um weitere Informationen zu erhalten, die systematisch für die Unterrichtsentwicklung genutzt werden können. Zugleich ist hiermit das Interesse nach einer Diskussion verknüpft, wie die Schülerresultate generell verwendet werden können. Dies lässt auf eine Unsicherheit bei den Lehrkräften schließen, indem sie selbst nur wenige Ideen hatten, wie eine effektive Nutzung der Rückmeldungen vorgenommen werden kann und sich daher Anregungen von ihren Kollegen erhofft haben. Dies war zugleich mit dem Bedürfnis nach einem Erfahrungsaustausch verbunden. Demnach sollten die an dem letzten Testdurchgang beteiligten Lehrkräfte die jetzigen über ihre Einschätzungen und Erfahrungen informieren beziehungsweise der Nutzungsprozess auch den nicht-beteiligten Lehrpersonen bewertend erläutert werden. Folglich erschien den Probanden eine Plattform zum Austausch solcher Erfahrungen als hilfreich. Ob diese Kommunikation institutionalisiert oder informell ablaufen soll, ist anhand der Interviewaussagen nicht eindeutig zu beantworten. Interessant erscheint hierbei insbesondere der Austausch mit nicht-beteiligten Lehrpersonen. Dieser Idee liegt die Zielsetzung zugrunde, das Konzept der Lernstandserhebungen stärker im Schulalltag zu verankern und das Kollegium damit vertraut zu machen.

Resümierend kann anhand der Aussagen zu den Kommunikationsbedürfnissen innerhalb des Kollegiums gefolgert werden, dass mindestens acht Probanden den Lernstandserhebungen mithilfe eines kollegialen Austauschs eine größere Bedeutung beimessen wollten. Das Potenzial einer wechselseitigen Kommunikation für die Qualität und Intensität des Nutzungsprozesses wurde erkannt. Das Konzept der Lernstandserhebungen wurde bei diesen Probanden folglich nicht generell abgelehnt. Vielmehr entwickelte sich bei ihnen das Bedürfnis, den Nutzen zu erhöhen und als eine Möglichkeit hierfür wurde der Austausch mit anderen Lehrpersonen angesehen.

Im Bereich der tatsächlich stattgefundenen inoffiziellen Gespräche wurden Besprechungen mit einzelnen Kollegen in Pausensituationen und Freistunden angeführt. Dieser Austausch erfolgte zufällig und verlief eher oberflächlich. Demgegenüber wurde in einem Fall von einem Erfahrungsaustausch mit einem teilnehmenden Kollegen des letzten Schuljahres berichtet. Der gleiche Proband (Lehrkraft 2) führte zudem gezielte Besprechungen mit der schulinternen Expertin für Lesekompetenz zum Zweck der Beurteilung der Deutschaufgaben sowie mit dem Fachsprecher durch. Ein Austausch über die Fächergrenzen hinweg konnte nicht festgestellt werden, weder bei den Lehrkräften noch bei den Mitgliedern der erweiterten Schulleitung.

Über den Inhalt der stattgefundenen inoffiziellen Gespräche geben folgende Interviewzitate Auskunft:

SL 2: Über die Qualität und die Ausgestaltung der Lernstandstests tauscht sich das Team aus. Das ist zu sehen. (SL 2, 00:30:53-6)

SL 3: Und es gibt aber immer wieder die Diskussion im Kollegium, zum Beispiel soll dafür geübt werden oder nicht, soll es dafür eine Note geben oder nicht? (SL 3, 00:14:53-2)

LK 6: Also man vergleicht zum einen mal die Klassenergebnisse und schaut sich auch dann noch mal - […], aber wir haben schon untereinander mal so gefragt: "Ja, bei welchen Aufgaben hat es denn gehakt oder welche Aufgaben sind denn generell nicht so gut ausgefallen?" Und das ist sehr unterschiedlich bei den einzelnen Klassen gewesen. (LK 6, 00:25:30-4)

SL 8: Und dann, eigentlich über die Aufgaben selbst haben wir weniger gesprochen, es sei denn, es waren halt irgendwelche missverständlichen Formulierungen bei diesen Aufgaben. Da hat man sich dann schon darüber unterhalten, aber ansonsten nicht so. (SL 8, 00:18:35-4)

LK 9: Also was wir […] gemacht haben, wir haben dann halt überlegt, also die beteiligt waren, wie können wir eventuell jetzt die Schüler und Schülerinnen, die gut abgeschnitten haben, wie können wie die irgendwie "belohnen" und haben da so einen Art Notenschlüssel festgelegt, der also dann auch Sinn macht. (LK 9, 00:22:14-4)

Im Fokus der Gespräche standen zum einen Ergebnisvergleiche zwischen den Parallelklassen hinsichtlich des Klassengesamtwerts und der Streuung. Zum anderen fanden organisatorische Absprachen sowie Diskussionen zur Intensität einer möglichen Vorbereitung der Schüler auf die Lernstandserhebung statt. Dabei fand eine inhaltliche Auseinandersetzung mit den Testinhalten, Aufgabenschwierigkeiten und Zielsetzungen statt. Ein kritisches Hinterfragen sowie ein Kommunizieren von negativ empfundenen Erfahrungen waren hierbei festzustellen. Daraus kann gefolgert werden, dass inoffizielle Gespräche oftmals bei negativen Assoziationen bezüglich der Testbedingungen geführt werden. Sie dienen dazu, die Erfahrung anderer Lehrkräfte mit der eigenen Einschätzung zu vergleichen und gegebenenfalls zu bestätigen. Zentrale Gesprächsaspekte nahmen insbesondere die Korrektur und Bewertung ein.

Lediglich ein Proband kommunizierte die Klassenergebnisse mit der Lehrkraft, welche die Klasse im folgenden Schuljahr unterrichten würde. Vor dem Hintergrund, dass der Ergebnisbericht erst Ende Mai erschienen ist und die Resultate durchaus auch für die anfängliche Diagnose im kommenden Schuljahr von Relevanz sein können, erscheint diese mangelnde Kommunikation überraschend. Dies spricht für einen geringen Stellenwert der Lernstandserhebungen in der Wahrnehmung der Lehrkräfte als ein nützliches Diagnoseinstrument, währenddessen die eigenen Beobachtungen als bedeutsamer eingeschätzt wurden.

Als Ursachen für eine nicht zustande gekommene oder nur geringfügige Kommunikation mit Kollegen wurden vielfältige Gründe von den Interviewprobanden angegeben.

SL 4: Die Lernstandserhebungen waren nicht mehr Thema, zumal das Schuljahr ja fast zu Ende war. (SL 4, Nachgespräch per E-Mail)

LK 2: Und ich habe zu wenige Chancen, im Schulalltag meine Ansicht mit Kollegen zu teilen, weil wir mit Arbeit überfordert sind. (LK 2, 00:55:56-0)

I: Auf welche Weise haben Sie sich mit Ihren Kollegen über diese Lernstandserhebungen ausgetauscht?

LK 5: Minimal, weil in Englisch Klasse 6 hat es keiner mehr gemacht. (LK 5, 00:14:58-9)

LK 1: Bei uns wäre das dann, also wenn wir das selber korrigieren, würde man natürlich auch nochmal mehr über die Ergebnisse sprechen. (LK 1, 00:29:30-4)

Zentrale Hindernisse für eine Kommunikation über die Lernstandserhebungen und deren Ergebnisse stellten einerseits die externen Rahmenbedingungen, wie der Zeitpunkt der Rückmeldung dar. Aufgrund der Zeitspanne zwischen der Testdurchführung und dem Erhalt des Ergebnisberichts ging das Interesse an den Lernstandserhebungen bei den Lehrkräften verloren. Andere Aspekte des Schulalltags rückten primär in den Fokus und minimierten die Gesprächsbedürfnisse. Die Probanden führten an, dass der Rückmeldungszeitpunkt mit dem Abitur kollidieren würde und die Arbeitsbelastung am Ende des Schuljahres zu hoch sei. Dies verdeutlicht erneut den geringen Stellenwert der Tests in der schulischen Arbeit, da anderen Aufgaben eine höhere Bedeutung zugemessen wurde. Die Lernstandserhebung wurde als eine zusätzliche Aufgabe betrachtet, deren kollegiale Auswertung zeitlich nur schwer zu bewältigen sei.

Ein weiteres Hindernis war die teilweise geringe Teilnahme des Kollegiums an der Leistungsmessung. Wenn lediglich ein bis zwei Lehrpersonen an den Tests partizipierten, entstand kein wechselseitiges Kommunikationsbedürfnis und die Lehrkräfte hatten nur geringfügig Möglichkeiten, die Ergebnisse kooperativ auszuwerten.

In einigen Schulen haben zudem externe Personen wie Studenten die Korrektur übernommen, so dass sich die Lehrkräfte nur geringfügig mit den Auswertungen beschäftigten. Wie aus der Aussage von Lehrkraft 1 ersichtlich wird, sank damit die Bereitschaft an einer Kommunikation über die Ergebnisse, da man selbst nur über wenige Informationen verfügte und somit keine solide Gesprächsgrundlage vorhanden war.

Neben dem inoffiziellen Austausch wurden die Lernstandserhebungen in acht der zwölf befragten Schulen in Fachkonferenzen thematisiert, wobei dies in den einzelnen Fächern nicht einheitlich gehandhabt wurde. Eine Betrachtung der Tests und der Ergebnisse in den Konferenzen bietet den Vorteil, dass alle Lehrkräfte eines Faches unabhängig von einer persönlichen Teilnahme mit dem Testkonzept und Aufgabenformaten konfrontiert werden. Der Umfang dieser Diskussion resultiert daraus, welche thematischen Aspekte auf der Fachkonferenz besprochen werden.

SL 1: Da werden die Aufgaben nochmal besprochen, wie die Resultate in den Klassen waren. (SL 1, 00:00:48-9)

SL 5: Also erst mal, ich habe informiert in der Deutsch-Fachschaft und die Kollegin, die in der 6 war, über die Gestaltung der Tests. Also auch, ich hab dann zum Beispiel berichtet […] mit dem Hörverstehen, dass es fünfzig Prozent war, dass eben Rechtschreibung und diese Dinge überhaupt nicht abgefragt werden. (SL 5, 00:14:59-7)

LK 3: Wir haben uns in Deutsch […] jetzt darauf geeinigt, das nicht zu tun. Also wir haben in der sechsten Klasse keine Lernstanderhebung Deutsch mehr. (LK 3, 00:14:32-9)

Im Fokus der Kommunikation innerhalb der Fachkonferenz standen sowohl die Ergebnisse der teilnehmenden Klassen, was meist in Berichtsform geschah, als auch eine Diskussion über die Inhalte, Schwierigkeiten und Bewertung der Lernstandserhebungen. Zudem wurde in einigen Schulen die Entscheidung über die weitere Teilnahme an der Leistungsmessung in der Fachkonferenz getroffen. Es setzt eine Auseinandersetzung mit den Zielsetzungen der Lernstandserhebungen sowie mit den Erfahrungen der Lehrkräfte voraus, um zu einer abschließenden Bewertung des Testkonzepts zu kommen. Die Intensität dieser Betrachtungen unterschied sich zwischen den einzelnen Schulen stark voneinander. Teilweise erfolgte lediglich eine oberflächliche Diskussion der Tests. Demgegenüber konnte in vier Schulen eine tiefgreifende Analyse des Testkonzepts und dessen Vereinbarkeit mit den eigenen pädagogischen und fachlichen Zielen festgestellt werden. Ein Zusammenhang zwischen einer intensiven Diskussion und einem positiven Entschluss zur weiteren zukünftigen Teilnahme konnte allerdings nicht beobachtet werden. Des Weiteren muss bezüglich dieser Form der Entscheidungsfindung kritisch angemerkt werden, dass auch Lehrkräfte über die weitere Teilnahme abstimmten, die bislang noch nicht an den Tests teilgenommen haben und nur auf Grundlage der berichteten Erfahrungen ihrer Kollegen ihr Votum abgeben konnten.

Konstruktive Diskussionen in Hinblick auf eine Nutzung der Lernstandserhebungen für die Schulentwicklung konnte lediglich bei der Lehrkraft 2 festgestellt werden:

LK 2: Und ich habe auch den konkreten Vorschlag gemacht, dass wir Aufgaben aus dieser Lernstandserhebung in Zukunft benutzen können für die Vergleichsarbeiten. […] Aber ich kann schwer abschätzen, in der Fachkonferenz fand ich ja eigentlich bis auf die eine Gegenmeinung oder sagen wir mal den kleinen Vorbehalt, fand ich da eigentlich kopfnickendes Zustimmen. Aber dabei ist es eben auch geblieben! (LK 2, 00:14:25-0)

Indem laut diesem Vorschlag die Aufgaben des Tests weitere Verwendung in den schulinternen Parallelarbeiten finden, soll eine nachhaltigere Nutzung angeregt und die Klassenarbeiten kompetenzorientierter gestaltet werden. Im folgenden Diskussionsprozess stellte der Proband letztlich eine befürwortende Haltung bei den Kollegen fest. Allerdings implizierte dies nicht zugleich ein Umsetzen des Vorschlages, wie aus dem Zitat ersichtlich wird,

da in der konkreten Situation der Parallelarbeit erneut auf herkömmliche Aufgaben zurückgegriffen wurde.

Die angeführten Hindernisse für einen geringfügigen inoffiziellen Austausch lassen sich auf die Kommunikation in der Fachschaft übertragen. Beispielsweise fand keine Konferenz unmittelbar nach dem Erhalt der Rückmeldung statt, sondern erst im kommenden Schuljahr, so dass die Lernstandserhebung keinen aktuellen Tagesordnungspunkt mehr darstellte. Weiterhin müssten auch Strukturen geschaffen werden, indem eine Thematisierung der Lernstandserhebungen von den Kollegen oder dem Fachsprecher initiiert wird. Dies setzt zugleich voraus, dass ein potentieller Nutzen der kollegialen Kommunikation zunächst einmal erkannt wird.

In Gesamtkonferenzen wurden die Lernstandserhebungen nur in einer der untersuchten Schulen thematisiert. Hierbei wurden vom Schulkoordinator die Testorganisation, die Zielsetzung und Aufgabenformate sowie erste Erfahrungen mit dem Testformat kommuniziert. Allerdings wurde in dieser Schule die Ansprache auch gezielt zur Steuerung eingesetzt. Im ersten Jahr der Teilnahme waren die Erfahrungen der Kollegen mit dem Testinstrument eher negativ. Daher wurde auf eine Vorstellung in der Gesamtkonferenz verzichtet, um Ablehnung und Demotivation zu vermeiden. Erst bei zunehmenden positiven Berichten entschloss sich die Schulleitung zu einer Thematisierung auf der Gesamtkonferenz, um die Unterstützung der Teilnahme im Kollegium zu festigen und zu erhöhen.

Mit Berücksichtigung dieser Ausnahme bildeten die Lernstandserhebungen keinen Gesprächsanlass in Gesamtkonferenzen. Dies war vor allem durch die geringe Anzahl der partizipierenden Lehrpersonen im Verhältnis zum Gesamtkollegium begründet.

Als Resümee kann angeführt werden, dass der kommunikative Austausch über die Lernstandserhebungen innerhalb des Kollegiums auf verschiedenen Intensitätsstufen verlief. Entsprechend der Auswertung zur Rezeption und Reflexion kann auch hier eine Kategorienbildung vorgenommen werden, die diese qualitativen Abstufungen verdeutlicht (vgl. Tabelle 11).

Kategorien zur Erfassung der Kommunikationsintensität bei den Lehrkräften		
Kategorie	**Beschreibung**	**Zugeordnete Personen**
keine Kommunikation	Es erfolgte keine Kommunikation über die Lernstandserhebungen im Kollegium.	LK 1, LK 10, LK 12
geringe Kommunikation	Die Kommunikation beschränkte sich auf einen Bericht und den Vergleich der Schülerergebnisse.	LK 5, LK 6, LK 7, LK 11
mittlere Kommunikation	Die Kommunikation umfasste eine Auswertung der Ergebnisse sowie eine Diskussion über die Testkonzeption.	LK 3, LK 4, LK 8
intensive Kommunikation	Die Kommunikation umfasste eine Auswertung der Ergebnisse sowie eine Diskussion über die Testkonzeption. Zudem wurden kollegiale Absprachen zur weiteren Nutzung der Tests getroffen.	LK 2, LK 9

Tabelle 11: Kategorien zur Erfassung der Kommunikationsintensität bei den Lehrkräften

Es erfolgte eine Einteilung in keine, geringe, mittlere und intensive Kommunikation. Während bei der geringen Kommunikationsstufe lediglich Informationen und Ergebnisse in Berichtsform ausgetauscht wurden, fand eine wechselseitige Diskussion erst auf der mittleren Intensitätsstufe statt. Die Diskussion bildete die Grundlage für gemeinsame Entscheidungen sowie zur Beurteilung des Testkonzepts. Die intensive Kommunikation zeichnete sich zusätzlich durch eine konstruktive Besprechung in Hinblick auf die Weiterarbeit mit den Lernstandserhebungen aus. Lediglich diese Stufe beförderte auf direkte Weise eine aktive Nutzung der Tests. In das Kategoriensystem wurden nur die befragten Lehrkräfte aufgeführt. Der Austausch mit Schulleitungsmitgliedern hat grundsätzlich andere Merkmale, da die Rechenschaftsfunktion hierbei eine stärkere Bedeutung einnimmt.

Es sollte angemerkt werden, dass sieben von zwölf Befragten keine Gespräche oder lediglich eine Vorstellung der Ergebnisse durchgeführt haben. Beide Kategorien haben keine Effekte auf die Nutzung der Lernstandserhebung für die Schulentwicklung. Daher kann gefolgert werden, dass eine konstruktive Diskussion innerhalb des Kollegiums in vielen Schulen noch nicht vorhanden ist.

Des Weiteren scheint eine intensive Reflexion der Tests nicht automatisch zu einer intensiven Kommunikation mit dem Kollegium zu führen. Dies ist darin begründet, dass ein kollegialer Austausch kommunikative Strukturen benötigt, die in den befragten Schulen bislang in unterschiedlicher Intensität ausgebildet waren (vgl. Abschnitt 10.2). Auch bedingen weitere Rahmenbedingungen, wie die Anzahl der teilnehmenden Lehrkräfte an der Leistungsmessung, massiv die Kommunikationsbedürfnisse nach einem solchen Austausch.

Kommunikation zwischen der teilnehmenden Lehrkraft und der Schulleitung sowie innerhalb der Schulleitung

Neben den Gesprächen innerhalb des Kollegiums kann eine Kommunikation mit Mitgliedern der Schulleitung stattfinden. In den zwölf untersuchten Schulen wurde dies lediglich bei der Hälfte vorgenommen. Als Gründe für die nicht vorhandene Kommunikation wurden einerseits Unwissenheit, inwiefern die Schulleitung überhaupt Interesse an den Lernstandserhebungen aufweist, und andererseits die geringe Bedeutung der Tests für die Schulleitung genannt. Letzteres spiegelte die Meinung wieder, dass die Ergebnisse lediglich den eigenen Unterricht beträfen und somit für die Schulentwicklung nicht von Belang seien. Folglich bestünde keine Notwendigkeit, die Schulleitung in die Auswertung der Lernstandserhebungen zu involvieren.

In den Schulen, bei denen eine Kommunikation mit der Schulleitung festgestellt werden konnte, gab es die folgenden Gesprächsaspekte:

LK 9: Also er hat jetzt im Prinzip nur so grob, wie die Klassen abgeschnitten haben. Also weniger jetzt inhaltlich, wo es fachlich hängt, sondern einfach das Gesamtergebnis der Klassen. […] Also es interessiert halt schon. (LK 9, 00:23:53-2)

SL 1: […] [Da] erbitte ich auch immer von den Fachkonferenzen eine offizielle Stellungnahme zu dem Wettbewerb. Also die diskutieren nicht nur, sondern ich will für die Schulleitung auch eine Rückmeldung von der Fachkonferenz in schriftlicher Form haben. (SL 1, 00:11:39-4)

SL 2: In diesem Jahr erstmalig […] wird sich die Schulleitung […] dransetzen und schauen: Wie kann ich abgleichen Vergleichsarbeit im Fach A mit Lernstandserhebung Fach A? […] Da muss man ein bisschen vorsichtig sein, nicht dass die Lehrkräfte das als Kontrolle und Scanning erleben und empfinden […]. (SL 2, 00:45:39-8)

SL 7: In der Schulleitung wurde das Thema noch einmal angesprochen. Hauptaspekt hier war die Frage, welchen Stellenwert die Lernstandserhebungen für den einzelnen Schüler haben, wenn Erfolg oder Misserfolg hier keine Konsequenzen für ihn haben. Inwiefern kann man also sicherstellen, dass die Schüler mit einer adäquaten Arbeitshaltung an den Lernstandserhebungen teilnehmen? (SL 7, Nachgespräch per E-Mail)

Die Schulleitungen nahmen primär die Gesamtergebnisse der Klassen wahr. Die Reflexion dieser Ergebnisse erfolgte mit sehr unterschiedlicher Intensität. Beispielsweise war für die Schulleitung 9 lediglich ein knapper Bericht über die Ergebnisse von Interesse. Eine weitere Auswertung nahm die Schulleitung nicht vor. Demgegenüber analysierte die Schulleitung 2 die Ergebnisse detaillierter, erfragte die Hintergründe zu den Leistungen und nahm weiterführende Vergleiche mit den Ergebnissen aus anderen Leistungsmessungen vor. Auf diese Weise erhielt sie differenziertere Informationen zu dem Leistungsvermögen der jeweiligen Lerngruppen. Diese Reflexion durch die Schulleitung war intrinsisch motiviert mit der Zielsetzung, die Ergebnisse weiterführend nutzen zu können.

Die Testkonzeption stellte ebenfalls einen Gesprächsgegenstand innerhalb der Kommunikationsebene mit der Schulleitung dar. Die Schulleitungen nahmen durchaus positive wie negative Bewertungen der Lehrkräfte zu den Lernstandserhebungen wahr und griffen dies zum Teil in eigenen Besprechungen auf. Die Diskussionen über die Testkonzeption fanden somit auf mehreren Ebenen statt, was eine höhere Relevanz der Lernstandserhebungen zur Folge hatte. Allerdings konnte dieser Aspekt der Reflexion nur bei den Schulleitungen 4 und 7 beobachtet werden.

Oftmals verlief die Kommunikation zwischen den Lehrkräften und den Schulleitungen nicht direkt, sondern es wurden die Fachkonferenzprotokolle an die Fachbereichsleiter weitergegeben, welche die Thematik anschließend in der Schulleitungskonferenz ansprechen konnten. Zwar erhielt die Schulleitung auf diese Wiese Einblick in die Diskussionsprozesse des Kollegiums, eine handhabbare Arbeitsgrundlage auf Schulleitungsebene stellte dies allerdings nicht dar.

Zusammenfassend kann festgestellt werden, dass die Kommunikation mit der Schulleitung in den meisten untersuchten Schulen einer Berichtsfunktion nahekam, die eine relativ oberflächliche Betrachtung der Gesamtergebnisse zur Folge hatte. Ein Grund hierfür ist in der Nicht-Beteiligung der Schulleitung an den Tests selbst zu sehen. Sie unterrichteten oftmals andere Fächer und Klassenstufen, so dass sie sich selbst nicht mit den Inhalten der Lernstandserhebungen auseinandergesetzt haben und nur geringfügige Kenntnisse zur Testkonzeption und Bewertungspraxis aufwiesen. Folglich fehlte die Gesprächsgrundlage für eine intensivere Kommunikation. Selbst bei Schulleitungsmitgliedern, die selbst an der Leistungsmessung teilgenommen hatten, war eine stark einseitige Ausrichtung auf ihr eigenes Fach festzustellen. Eine fächerübergreifende Reflexion war bei keinem der Probanden zu erkennen.

Kommunikation zwischen der teilnehmenden Lehrkraft und ihren Schülern

Als weitere Kommunikationspartner sind die Schüler anzuführen. Da in den Lernstandserhebungen ihre Kompetenzen ermittelt werden, weisen die Lernenden ein natürliches Informationsbedürfnis nach einem Feedback zu ihren Resultaten auf. Somit stellt die Kommunikation der Schülerergebnisse ein bedeutsames Element innerhalb des Nutzungsprozesses dar.

Ein Großteil der befragten Probanden besprach die Lernstandserhebungen in der Großgruppe der Klassengemeinschaft. Dies wurde durch das Inhaltskonzept der Rückmeldungen gefördert, da aggregierte Durchschnittswerte der Gesamtklasse ausgewiesen wurden. Der einzelne Schüler konnte seinen individuellen Punktwert nur dann einsehen, wenn er seinen

zugehörigen Verschlüsselungscode kannte. Es ist zu hinterfragen, welchen Erkenntniswert eine solche Besprechung in der Großgruppe für den individuellen Lernenden hat. Im Klassenwert kann er seine eigene Leistung in keiner Weise verorten.

Lediglich vier teilnehmende Lehrkräfte führten persönliche Gespräche zu den Ergebnissen mit den Schülern. Auf diese Weise ermöglichten sie ihnen, ihre Stärken und Schwächen in den einzelnen Bereichen des Tests zu erkennen. Diese Hinweise sind nicht nur für die Lehrkraft als Diagnoseinformationen von Bedeutung, sondern können dem Schüler als Orientierung für den weiteren Lernprozess dienen. Voraussetzung bildet jedoch eine intensiv vorgenommene Reflexion der Rückmeldung durch die Lehrperson.

Bei der Erhebung konnten insgesamt drei Kommunikationsaspekte zwischen Lehrenden und Lernenden festgestellt werden: die Bewertung des Tests, die Auswertung der Ergebnisse und die Besprechung der Aufgaben. Im Rahmen des ersten Bereichs, der Bewertung der Tests, fassen die folgenden Interviewpassagen die wesentlichen Gesprächsinhalte zusammen:

> *LK 1: Ja, ich habe eigentlich vor allem besprochen, inwieweit das jetzt zu viel Druck war, weil die waren schon ausgepowert. (LK 1, 00:30:27-0)*
>
> *LK 3: Ja, wir haben darüber gesprochen, wie es Ihnen ging. Ja also, das musste man einfach tun, weil so viele traurige Gesichter da einfach saßen, die dann gedacht haben, sie hätten versagt. (LK 3, 00:07:01-3)*
>
> *LK 6: Ansonsten haben meine Schüler mir rückgemeldet direkt nach dem Test, dass sie es insgesamt relativ leicht fanden. Also gerade so was den Anfang der Hörverstehensübungen betrifft […]. (LK 6, 00:17:28-3)*

Die drei Lehrkräfte besprachen somit mit ihren Schülern, wie sie die Lernstandserhebung hinsichtlich Schwierigkeit, getesteter Inhaltsbereiche und äußerer Rahmenbedingungen empfanden. Diese Kommunikation verlief stets unmittelbar im Anschluss an den Test. Die Lehrpersonen erhielten auf diese Weise ein Stimmungsbild, was ihre eigene Bewertung der Lernstandserhebungen beeinflusste. Beispielsweise ist bei Lehrkraft 1 und 3 insgesamt eine schlechtere Beurteilung des Testinstruments zu konstatieren. Indem ihre jeweiligen Lerngruppen negative Assoziationen mit der Leistungsmessung formulierten, wirkte sich dies auf die Grundhaltung der Lehrpersonen zu dem Test aus. Diese Einstellung wurde im Verlauf des Nutzungsprozesses durch weitere Erfahrungen verstärkt. Die Kommunikation mit den Schülern war zwar nicht die einzige Ursache für die Ablehnung der Lernstandserhebung, beförderte sie jedoch immens.

Im Kontext der Ergebnisbesprechung geben folgende Interviewzitate für die Kommunikation im Unterricht Aufschluss:

LK 5: Ich habe allgemein gesagt, soundso viel Prozent und […] in welchem Feld sie so lagen, wer da im Spitzenfeld war und wer eher hintendran war. (LK 5, 00:07:40-8)

LK 7: Ich habe nur so allgemein dann noch etwas dazu gesagt und dann, wie gesagt, eben individuell, in welchem Bereich das liegt notentechnisch gesehen, habe ich das so ungefähr umgerechnet mit den Punkten und Prozenten. (LK 7, 00:18:08-5)

LK 12: Und es hatte dann jeder Schüler sein Heft zurückbekommen, wo die einzelne Korrektur für den Schüler drin war, wo er also sehen konnte: "Was hatte ich richtig, was nicht?" […] Und ich habe dann auch für zwei Tage diese grafischen Ergebnisse im Klassenrahmen ausgehängt. Weil sie es einfach interessiert hat! (LK 12 + SL 12, 00:24:08-6)

Ein Großteil der befragten Personen gab an, eher eine informierende Auswertung in Form eines Berichts in ihrer Klasse vorgenommen zu haben. Dies implizierte eine Vorstellung und Projizierung der Ergebnisdiagramme, so dass die Schüler einen Gesamtüberblick über ihr Abschneiden im Test gewinnen konnten. Einige gaben zudem die Aufgabenhefte zurück. Die Schüler erhielten damit die Möglichkeit, sich mit ihrer individuellen Leistung auseinanderzusetzen. Da jedoch keine ausführliche Korrektur im Schülerheft von der Lehrkraft vorgenommen wurde, erhielten die Schüler zum Großteil unkorrigierte Aufgabenhefte ohne individuelle Bemerkungen zurück, was die persönliche Auswertung massiv erschwerte. Zudem haben die Schüler keine Kenntnis von den Bewertungsgrundlagen und Punktverteilungen. Lediglich eine Lehrkraft erläuterte den Schülern die Korrekturprinzipien.

Interessant ist des Weiteren das von der Lehrkraft 12 angesprochene Aushängen der Resultate. Somit erhielten die Schüler Zeit, sich in Ruhe mit den Ergebnissen auseinanderzusetzen. Dies erhöhte die Wertigkeit der Lernstandserhebungen, da den Ergebnissen ein zentraler Ort im Klassenraum zur Verfügung gestellt wurde. Eine zusätzliche Möglichkeit, die Auswertung für die Schüler zu erleichtern, sprach Lehrkraft 7 an, indem sie die individuellen Punktwerte in Notenwerte umrechnete. Die Notenskala stellte für die Lernenden eine für sie bekannte Bezugsgröße dar, so dass sie die eigene Leistung gezielter beurteilen konnten. Problematisch ist jedoch die willkürliche Festsetzung dieses Maßstabes in Bezug auf die Punktwerte des Tests durch die Lehrkraft. Aufgrund der testtheoretischen Hintergründe kann der herkömmliche Bewertungsmaßstab nicht automatisch auf die Lernstandserhebungen angewandt werden.

Die Aufgabenbesprechung stellte bei fünf Lehrkräften einen Gesprächsaspekt dar.

LK 8: Ja, also ich habe einzelne Übungen nochmal so mit ihnen gemacht und auch so besprochen, wo sie da Schwierigkeiten hatten. (LK 8, 00:15:49-3)

Hierbei wurden einzelne Aufgaben diskutiert und die Lösungen erläutert. Stets wurden nur diejenigen Items thematisiert, bei denen die Schüler Probleme aufgewiesen hatten und die

Klasse im Vergleich zum Landesmittelwert unterdurchschnittlich abgeschnitten hatte. Diese Form der Auswertung ist vergleichbar mit der üblichen Nachbesprechung und Berichtigung einer Klassenarbeit. Des Weiteren wurde eine Verbindung zwischen der jeweiligen Aufgabenanforderung und dem Klassenergebnis hergestellt. Der einzelne Schüler erhielt zudem eine informative Grundlage für die eigene Reflexion seiner Leistung.

Insgesamt nahm bei fast allen befragten Personen die Besprechung der Ergebnisse mit den Schülern nicht mehr als eine Unterrichtsstunde in Anspruch. Lediglich zwei Lehrkräfte führten intensivere Auswertungen durch, deren Ausmaße im Folgenden kurz vorgestellt werden.

> *LK 2: Dann habe ich aber die wesentlichen Aufgaben […], die habe ich dann mit den Kindern nochmals gelöst und zwar in Gruppenarbeit. […] Und als jetzt […] das Ergebnis da war, habe ich meinen Ergebnisbericht […] den Schülern als Folie projiziert und wir haben über diese Aufgaben erneut gesprochen. Und jetzt seit letzte Woche Donnerstag […] arbeiten wir nochmal die Aufgaben durch, schwerpunktmäßig diese Aufgabe "Anna-Sophia" und haben mit den Ergebnissen, wie sie in den Diagrammen sichtbar werden, überlegt: "Wie kommt es, dass so viele von euch bei bestimmten Teilaufgaben versagt haben, während sie bei anderen Teilaufgaben gute Ergebnisse erzielt haben?" […] Ich habe dagegen gesetzt eine typische beurteilende Fragestellung als Analyse zu einer solchen Kurzgeschichte. Und wir haben jetzt eigentlich heute in der Stunde, mit meiner massiven Hilfe muss ich dazu sagen, sind wir zu einem Ergebnis gekommen, dass solche übergreifenden Fragestellungen eigentlich sinnvoller sind. (LK 2, 00:04:34-5)*

Lehrkraft 2 hat bedeutsame Aufgaben der Lernstandserhebung in Gruppenarbeit von den Schülern nochmals lösen lassen. Zudem wurden einzelne Items nach Erhalt des Ergebnisberichts vor dem Hintergrund der durchschnittlichen Klassenleistung und der Bewertungsgrundlagen kritisch reflektiert. Zielsetzung war das Erkennen der zugrundeliegenden Ziele einer Aufgabe anhand eines Vergleichs zwischen Lern- und Testaufgaben und die Bewertung deren Sinnhaftigkeit. Die Lehrkraft konzipierte auf Basis der Lernstandserhebung eine Unterrichtsreihe im Umfang von sechs Unterrichtsstunden, bei der neben der ausführlichen Aufgaben- und Ergebnisbesprechung die Reflexionskompetenz der Schüler gefördert wurde.

Ähnlich intensiv nahm die Lehrkraft 10 die Auswertung in ihrer Klasse vor.

> *LK 10: Und dann ist für mich das Wichtigste, dass ich in so einer Art Workshop jedem Einzelnen seine Textproduktion vorlege mitsamt der Korrekturen, die ich angefertigt habe. Ich werde die Korrekturzeichen, die ich verwendet habe, erklären, denn die sind den Schülern unbekannt, weil eben Textproduktion neu ist für sie. Und werde ihnen die Bedeutung der Textproduktion nochmal nachhaltig erklären […]. Naja, und dann werde ich mich mit jedem Einzelnen zusammensetzen und werde das in Ruhe besprechen […] und werde von jedem Einzelnen verlangen als Hausaufgabe, dass der geschriebene Text sozusagen berichtigt wird. (LK 10, 00:10:46-2)*

Lehrkraft 10 führte die Lernstandserhebung in einer achten Klasse im Fach Deutsch durch. Zentraler Testinhalt war der bedeutsame Kompetenzbereich der Textproduktion, welcher zuvor im Unterricht noch nicht behandelt worden war. Die Lehrkraft 10 nahm die Lernstandserhebung zum Anlass, die Textproduktion intensiv einzuführen. Aus diesem Grund korrigierte sie die Schülerarbeiten detailliert und besprach sie mit den Lernenden ausführlich, so dass diese aus ihren eigenen Texten heraus die Merkmale einer Textproduktion verstehen konnten. Zugleich wurden sie erstmalig mit den zugehörigen Korrekturzeichen vertraut gemacht, welche bei späteren Arbeiten ebenfalls Verwendung finden würden. Die Besprechung dauerte insgesamt vier Unterrichtsstunden.

Eine solche intensivere Kommunikation konnte nur bei den Lehrkräften 2 und 10 festgestellt werden. Bei den anderen erfolgte eine Auswertung der Lernstandserhebungen oberflächlicher bis gar nicht. Um einen Überblick über die Intensität der Kommunikation mit den Schülern zu ermöglichen, erfolgt erneut eine Kategorisierung mit vier Abstufungen (vgl. Tabelle 12).

Kategorien zur Erfassung der Kommunikationsintensität mit den Schülern		
Kategorie	**Beschreibung**	**Zugeordnete Personen**
keine Kommunikation	Die Lernstandserhebung wurde nicht reflektiert oder es erfolgte lediglich ein Austausch von Wahrnehmungen unmittelbar nach dem Test.	LK 1, LK 3, LK 11, SL 4, SL 12
geringe Kommunikation	Die Schüler wurden über die Ergebnisse der Klasse informiert und/ oder die Testhefte wurden zurückgegeben.	LK 4, LK 5, LK 7, SL 3, SL 8, SL 9
mittlere Kommunikation	Die Ergebnisse wurden individuell oder in der Großgruppe diskutiert und/oder einzelne Aufgaben wurden besprochen.	LK 6, LK 8, LK 9, LK 12, SL 1, SL 5
intensive Kommunikation	Es erfolgte eine intensive Auswertung, indem die Lernstandserhebung die Grundlage für eine Unterrichtseinheit bildete.	LK 2, LK 10

Tabelle 12: Kategorien zur Erfassung der Auswertungsintensität mit den Schülern

Das Kategoriensystem konzentriert sich auf die Auswertung der Lernstandserhebungen mit den Schülern. Unmittelbar im Anschluss an den Test vorgenommene Gespräche über die Wahrnehmung des Zeitumfangs etc. wurden der Kategorie „keine Auswertung“ zugeordnet. Ein solcher Austausch liefert zwar auch Anhaltspunkte, ermöglicht den Schülern jedoch keinen konstruktiven Erkenntnisgewinn in Hinblick auf ihren weiteren Lernprozess. Die „geringe Auswertung“ erfolgt lediglich als Informationsweitergabe in Berichtsform, während bei der „mittleren Auswertung“ die Ergebnisse bereits diskutiert und/ oder einzelne Aufgaben vertieft analysiert werden. Die „intensive Auswertung“ besteht in der Verwendung der

Lernstandserhebung und Rückmeldungen, um weiterführende Kompetenzbereiche zu fördern, wie an den Beispielen der Lehrkräfte 2 und 10 ersichtlich wurde.

Bei Betrachtung der Verteilung der befragten Probanden wird ersichtlich, dass zehn von 19 teilnehmenden Lehrpersonen keine bis lediglich eine geringe Auswertung mit den Schülern vornahmen. Es ist zweifelhaft, ob deren Schüler letztlich von den Lernstandserhebungen persönlich profitieren konnten, wenn sie keine Rückmeldung zu ihrer Leistung bekamen beziehungsweise dem Test keinen Stellenwert im Unterricht eingeräumt wurde. Als Hindernisse für eine ausführliche Auswertung der Leistungsmessung wurden folgende Aspekte von den Befragten angeführt:

> *LK 8: Also ich hab dann auch nicht so viel Zeit damit verbracht, weil ich fand, es war nicht unbedingt nötig. (LK 8, 00:18:00-3)*
>
> *SL 12: Wir sollten ja erst warten, bis diese allgemeinen Daten da sind und dann könne man den Schülern die Lernstandserhebung zurückgeben. […] Es ist halt jetzt ein bisschen viel Zeit vergangen […]. (LK 12 + SL 12, 00:22:30-7)*
>
> *LK 4: Erstens mal kriegen die Schüler ja durch die Art der Korrektur nicht die Rückmeldung, die sie eigentlich brauchen, um die Aufgaben auch zu verstehen bzw. ihre Leistung zu verstehen. Wir sind da gezwungen durch die schnelle Bewertung […] festzustellen, ob die Aufgabe erfüllt wurde oder nicht. […] Da müsste man nochmal auf die Aufgabe eingehen und den Schülern nochmal speziell erklären, hier ist die Leistung nicht erreicht, weil-. (LK 4, 00:09:11-1)*
>
> *LK 6: Andererseits muss man dann auch sagen, […] dass Rechtschreibung und Zeichensetzung oder so die sprachlichen Aspekte des Faches Deutsch kaum bis gar nicht berücksichtigt werden. Und die Vermittlung dessen dann eben auch an die Schüler, also erklären, warum spielt das denn hier in dem Fall keine Rolle und sonst wird es in den Arbeiten aber benotet. Das ist einfach so ein Balanceakt […]. (LK 6, 00:32:40-2)*

Als ein triviales Hindernis führte Lehrkraft 8 die fehlende Notwendigkeit einer Besprechung mit den Schülern an. Wenn die Lehrperson selbst nur einen geringen bis keinen Nutzen in den Lernstandserhebungen sieht, wird sie ihre Unterrichtszeit nicht erneut darauf verwenden wollen. Allerdings wird auf diese Weise das Informationsbedürfnis der Lernenden missachtet, welche eine Rückmeldung ihrer Leistung erhalten möchten.

Die weiteren kommunizierten Hindernisse sind in den Rahmenbedingungen der Tests begründet. Erneut wurde die erhebliche Zeitspanne zwischen Test und Erhalt der Rückmeldung angeführt, so dass die Leistungsmessung sowohl für die Lehrkräfte als auch für die Schüler nicht mehr präsent gewesen sei. Des Weiteren wurden die Korrekturprinzipien als Hindernisse benannt. Zum einen böte die Form der Korrektur keine erkenntnisreiche Rückmeldung für die Schüler, so dass die Lehrkraft zusätzliche Diagnosen vornehmen und kommunizieren müsste. Zum anderen würden alltägliche Bewertungsrichtlinien, wie das Anstreichen von Rechtschreibfehlern oder die Vergabe von Teilpunkten, bei den Lern-

standserhebungen außer Kraft gesetzt. Dies sei problematisch zu besprechen, da die Schüler hierdurch verwirrt würden und ein Hinterfragen des herkömmlichen Bewertungsmaßstabes möglich sei.

Kommunikation zwischen den teilnehmenden Lehrkräften und den Eltern

Ein weiterer möglicher Gesprächspartner ist die Elternschaft. Die Eltern wünschen ebenfalls Informationen und eine Rückmeldung zum Leistungsstand ihres Kindes, so dass die Ergebnisse der Lernstandserhebung als Gesprächsgrundlage dienen können. Von den befragten 19 Personen, welche die Tests in ihrer Klasse durchführten, kommunizierten zehn Lehrkräfte mit der Elternschaft über die Lernstandserhebung. Hinweise zu Intensität, Form und den thematisierten Aspekten liefern die folgenden Aussagen:

LK 8: Die Kurven habe ich Ihnen auf Folie gemacht, dass sie es so ein bisschen sehen konnten. Solche Sachen halt, also was so anonym möglich war. Ich konnte ja nicht über Einzelne da reden im Elternabend. Und damit waren die eigentlich zufrieden. (LK 8, 00:33:40-9)

SL 4: Also ich habe denen einen Brief geschrieben und habe auch Auszüge den Eltern kopiert aus diesem Ergebnisbericht da und habe ihnen natürlich gesagt, ich bin auch gerne bereit, […] ihnen das zu erläutern. Aber die Zeit war viel zu lang dazwischen. (SL 4, 00:37:34-9)

SL 5: Und sie haben da eigentlich auch nichts dagegen gehabt, dass ich das jetzt als mündliche Note werte. Das ist ein bisschen heikel, also da hätten die Eltern natürlich sagen können: "Nein, das dürfen Sie nicht und so weiter." (SL 5, 00:26:21-3)

LK 7: Ich habe das auch unter die folgende Englischarbeit drunter geschrieben […]: "Lernstandserhebung wäre Note X". Ja also ich habe schon so eine kurze schriftliche Information an die Eltern gegeben. Aber es hat jetzt keiner von den Eltern bei mir das Gespräch gesucht um das nochmal zu diskutieren. (LK 7, 00:36:41-2)

LK 4: Und ich bin selbst zu dem Schluss gekommen, dass es nicht sinnvoll ist. Man könnte die Klasse als solches darstellen […], aber die Eltern interessiert doch eher die individuelle Rückmeldung bezüglich des Kindes. (LK 4, 00:32:10-8)

LK 6: Mit den Eltern von dem hörgeschädigten Schüler habe ich natürlich im Nachgang telefoniert und mit ihnen das besprochen, was mir eben bei ihm aufgefallen ist. (LK 6, 00:28:54-7)

Aus den Interviewpassagen wird ersichtlich, dass verschiedene Formen der Kommunikation eingesetzt wurden. Zum einen wurden die Ergebnisse auf einem Elternabend besprochen und zum anderen wurden die Eltern schriftlich über einen Brief informiert. Beide Möglichkeiten sind als eine relativ einseitige Informationsweitergabe zu betrachten, da keine wechselseitige Diskussion entstanden ist. Lediglich der Schulleitungsvertreter 5 deutete an, mit der Elternschaft die Verwendung der Ergebnisse im Rahmen einer mündlichen Note besprochen und abgeklärt zu haben. Des Weiteren wurden bis auf eine Ausnahme lediglich

die Durchschnittswerte der Gesamtklasse thematisiert. Zu begründen ist dies mit dem Gebot der Anonymität jedes einzelnen Schülers. Zum anderen bieten die Rückmeldungen keine individuellen diagnostischen Hinweise, welche kommuniziert werden könnten. Hierfür wäre wiederum eine ausführlichere Auswertung der Schülerarbeiten durch die Lehrkraft erforderlich. Dies stellte zugleich einen von Lehrkraft 4 geäußerten Hindernisfaktor für eine Kommunikation mit den Eltern dar: Diese würden individuelle, auf ihr Kind bezogene Informationen benötigen, welche die standardisierte Rückmeldung jedoch nicht zur Verfügung stellt.

In einem Fall besprach Lehrkraft 6 die Ergebnisse eines einzelnen Schülers mit dessen Eltern. Dieses Kind wies eine Lernbeeinträchtigung auf. Daher war im konkreten Fall eine intensivere Lehrer-Eltern-Kommunikation bereits etabliert. Oftmals boten die Lehrpersonen weiterführende Gespräche bei Bedarf an, welche allerdings von keinem Elternteil in Anspruch genommen wurde. Folglich schien kein besonders großes Informationsbedürfnis auf Seiten der Eltern vorzuliegen beziehungsweise es wurde dem Abschneiden bei der Lernstandserhebung keine große Bedeutung beigemessen.

Kommunikation zwischen schulischen Akteuren und außerschulischen Personen bzw. Institutionen

Die Etablierung von Kommunikationsstrukturen zwischen der Schule und außerschulischen Personen und Institutionen wie dem Schulamt sind für die Lernstandserhebungen vom LSA nicht intendiert. Dennoch konnten bei den Interviews vereinzelt stattgefundene Gespräche ermittelt werden:

> *SL 1: Schulamt hat im letzten Jahr nicht nachgefragt. Wir haben nur unsererseits unaufgefordert unsere Meinung zu den Tests genannt. (SL 1, 00:29:55-8)*
>
> *SL 6: Unsere Schule schnitt allerdings bei diesen Lernstandserhebungen gut ab […], sodass wir dies an die örtliche Presse gaben und darüber auch ein Artikel erschien. (SL 6, Nachgespräch per E-Mail)*
>
> *LK 2: Also als die Schüler ihre Gruppenberichte über diese Aufgaben abgeliefert haben, waren die Referendare mitbeteiligt […]. Also von da aus habe ich diese Sache damals in die Ausbildung gebracht […].(LK 2, 00:04:34-5)*

Die Schulen 1 und 6 gaben aus intrinsischer Motivation heraus die Ergebnisse nach außen weiter. Die Maßnahmen beider Schulen verstärkten die Rechenschaftsfunktion der Lernstandserhebungen. Schule 6 legte zudem offensiv Wert auf die Veröffentlichung der Ergebnisse, um sich auf diese Weise im direkten Vergleich mit Nachbarschulen zu profilieren. Dies steht zwar den Forderungen nach Anonymität und dem Vermeiden eines öffentlichen Rankings vehement entgegen, doch der Schule stand das Verfolgen der eigenen Zielsetzungen im Vordergrund.

Eine andere Form der außerschulischen Kommunikation stellte die Thematisierung im Kontext der Lehrerausbildung im Studienseminar dar, wie sie Lehrkraft 2 beschrieb. Sie selbst wie auch einige Lehrkräfte im Vorbereitungsdienst führten die Lernstandserhebungen in ihren Klassen dadurch, so dass es innerhalb der Seminargruppe als aktuelles Thema analysiert wurde.

Neben diesen aufgezählten außerschulischen Personen und Institutionen ist ebenso das LSA zu erwähnen. Insgesamt nahmen acht der neunzehn befragten Personen, die an den Lernstandserhebungen teilgenommen hatten, Kontakt mit dem LSA auf beziehungsweise beantworteten den Online-Feedbackbogen.

LK 8: Also ich hatte das Gefühl, die haben es sehr von sich fern gehalten, dass man ja nicht so viel schreiben konnte. Also wo ich es konnte, habe ich das genutzt. […] Aber ich fand, da hätte man ein bisschen mehr Raum kriegen können dafür. Also ich war mir nicht so sicher, ob die wirklich das wollten, dass man das kritisiert, was sie da machen. (LK 8, 00:37:40-8)

SL 4: Und natürlich hat jeder Kollege oder viele Kollegen haben, ich auch, diesen kurzen Kommentar da in der Internet-Befragung auch ausgefüllt. […] Aber der hilft gar nicht, weil die nämlich davon ausgehen, dass das, wie da getestet wird, dass das erst mal Konsens findet. […] Also es wird nicht gefragt, sind die Aufgaben angemessen? Ist die Zeit und so weiter? Nichts! […] Das finde ich ein zweifelhaftes Verfahren. (SL 4, 00:13:42-6)

LK 3: Wir haben etwas geschrieben. Wir haben uns nochmal zusammengesetzt und haben Kritikpunkte gesammelt und dahin geschrieben. Die Fachbereichsleiterin hat das gemacht. (LK 3, 00:30:11-8)

LK 2: Dann habe ich Anfang der Ferien meine Unterlagen, die ich dazu hatte, also die Ergebnisse der Erarbeitung mit den Schülern, die Rückmeldung durch die Referendare, aber auch eigene Kritik an den Aufgabenstellungen, […] an das [Institut] geschickt mit einem ziemlich ausführlichen Begleitschreiben. Und ich bekam dann Ende der Sommerferien eine Antwort. […] Da hat eine Dame […] mir geantwortet und sich einmal bedankt für den Aufwand, den ich betrieben habe, dann zu einzelnen meiner Kritikpunkte sich knapp geäußert. […] Also mit anderen Worten, die betreffende Dame hat sich gewundert, wie viel Energie und Aufwand ich da hineingesteckt habe. Sie hat das natürlich lobend vermerkt. Aber es klang eher so, als ob das für sie ungewohnt sei. (LK 2, Nachgespräch, 00:04:15-8)

Bezüglich des Feedbackbogens konnte durchgängig immense Kritik wahrgenommen werden. Die Lehrkräfte hatten sich erhofft, auf diese Weise eine ernstgemeinte Rückmeldung dem Institut für Qualitätsentwicklung (Vorgängerinstitut zur LSA) übermitteln zu können, so dass auf dieser Grundlage die Lernstandserhebungen weiterentwickelt und verbessert werden können. Allerdings wurde diese Erwartung nicht erfüllt, denn der Feedbackbogen ist vorrangig auf die Erfassung des Nutzungsprozesses bei den Lehrkräften ausgerichtet. Die Lehrkräfte erachteten eine Möglichkeit zur direkten Kommunikation mit dem Institut als notwendig und wollten in die inhaltliche Gestaltung der Tests einbezogen werden. Statt-

dessen entstand Frustration, da der Eindruck vermittelt wurde, dass Diskussionen vermieden würden und kritische Stimmen nicht erwünscht seien. Auf diese Weise vergrößerte sich die Distanz zwischen den schulischen Akteuren und dem Institut, was die Bewertung der Tests negativ beeinflusste.

Als Alternative zu diesem Feedbackbogen formulierten die Lehrkräfte 2 und 3 ihre Einschätzungen schriftlich in Form eines Briefes. Es entstand daraus eine wechselseitige Kommunikation lediglich bei Lehrperson 2, welche ein Antwortschreiben erhielt. Die von der Lehrkraft geäußerten Kritikpunkte wurden mit testtheoretischen Gründen gerechtfertigt, was für sie keine zufriedenstellende Antwort darstellte. Wenn eine unsinnige Aufgabenstellung aufgrund der Testentwicklung nicht anders konzipierbar sei, ändere dies nichts an der eigenen Beurteilung und der Tatsache, die Schüler mit dieser Aufgabe konfrontieren zu müssen.

Die Kommunikation mit dem Institut wurde zusammenfassend von der Mehrheit der Befragten als negativ, stark einseitig und der Feedback-Bogen als mangelhaft konstruiert wahrgenommen.

9.3 Aktion

Nach dem Wirkungsmodell von Helmke (vgl. Helmke A. , 2004, S. 100) schließt sich an die Phasen der Rezeption und Reflexion die Aktion an, in der Maßnahmen entwickelt und umgesetzt werden. Folglich ist dieser Teil des Nutzungsprozesses von besonderer Bedeutung für die Anregung von Schulentwicklung durch die Lernstandserhebungen. Mithilfe der Maßnahmenkonzeption werden tatsächliche oder unbewusste Veränderungen angestoßen, welche zur Qualitätsentwicklung beitragen sollen. Aus diesem Grund stellte die Aktion einen zentralen Auswertungsgegenstand der Erhebung dar. Anknüpfend an die theoretischen Ausführungen in Abschnitt 5.1 erfolgt eine Unterteilung der Ergebnissdarstellung in die drei Bereiche der Schulentwicklung: der Unterrichts-, Personal- und Organisationsentwicklung.

9.3.1 Unterrichtsentwicklung

Maßnahmen im Kontext der Unterrichtsentwicklung betreffen in der Regel den eigenen Unterricht. Die Lehrkraft zieht aus den Ergebnissen ihrer Reflexion Konsequenzen für die Weiterarbeit in ihrer Klasse. Die Probanden wurden in diesem Zusammenhang befragt, welche Anregungen sie aus den Lernstandserhebungen und den Schülerwerten für den

Unterricht ziehen konnten und inwiefern sie diese bereits praktisch umgesetzt haben. Dabei kristallisierten sich vier Bereiche heraus, denen die geplanten Maßnahmen zuzuordnen waren:

- Bewertung der Schülerergebnisse,
- Anregungen durch Aufgabenformate,
- individuelle Förderung sowie
- Anregungen durch Testinhalte.

Im Folgenden werden die Ergebnisse zu diesen Bereichen dargelegt.

Bewertung der Schülerergebnisse

Die Schülerleistungen in den Lernstandserhebungen sollen laut Aussage des zuständigen Instituts LSA nicht benotet werden, da sie einen größeren Zeitraum der Schullaufbahn testen, als es für schriftliche Leistungsüberprüfungen gestattet ist. Dennoch berichteten sechs befragte Personen, welche die Lernstandserhebungen in ihrer Klasse durchgeführt hatten, von einer Bewertung der Schülerleistungen, wie in den folgenden Auszügen geschildert wird:

> *LK 7: Deswegen habe ich die Lernstandserhebung ja auch, wie gesagt, schon in meine Englischnoten mit einfließen lassen. […] Denn in dem Fall musste ich ja arbeiten und die Schüler haben ja auch gearbeitet in dem Sinne. Warum soll das nicht auch gewertet werden? […] Und wenn man ihnen im Vorfeld sagt, dass man sich das schon genauer anschaut […], gehen sie auch mit größerer Ernsthaftigkeit daran. (LK 7, 00:35:32-4)*

> *SL 4: Wir hatten überlegt, wollen wir daraus eine Vergleichsarbeit machen? […] Wir hatten gehofft, das spart uns vielleicht eine Klassenarbeit auch mal anders und dann stellt sich heraus, das geht überhaupt nicht. (SL 4, 00:11:37-5)*

> *LK 9: […] [Wir] haben dann halt überlegt, also die beteiligt waren, wie können wir eventuell jetzt die Schüler und Schülerinnen, die gut abgeschnitten haben, […] irgendwie "belohnen" und haben da so einen Art Notenschlüssel festgelegt, der also dann auch Sinn macht. Dass es eben nicht nur Einsen gibt, aber dass es eben so im Schnitt die Noten sind, die die Schüler schreiben. (LK 9, 00:22:14-4)*

In den beobachteten Fällen wurden die Lernstandserhebungen als eine Teilnote in dem Bereich der sonstigen Leistungen berücksichtigt. Lehrkraft 7 beschrieb deutlich die Gründe für ein solches Vorgehen: Zum einen fungierte die Bewertung als Motivation, damit die Schüler möglichst konzentriert und ehrgeizig die Lernstandserhebung absolvierten. Zum anderen sollte die Note die Arbeit der Schüler wie auch den Korrekturaufwand der Lehrperson honorieren und kompensieren. Dies macht deutlich, wie stark schulische Prozesse auf die Notengebung fokussiert sind. Lediglich mit einer Benotung sei es möglich, die Leistung zu würdigen. Letzteres wurde durch die Aussage von Schulleitung 4 bestätigt, indem durch

die Wertung der Testergebnisse als eine Vergleichsarbeit eine zusätzliche Korrektur für die Lehrpersonen vermieden werden sollte, was rechtlich jedoch nicht möglich ist. Die Schülerleistungen wurden daraufhin zwar nicht als Klassenarbeit gewertet, aber die festgelegte Vergleichsarbeit wurde dennoch nicht geschrieben.

Drei Lehrkräfte schwächten die Entscheidung zur Benotung der Testresultate ab, indem lediglich diejenigen Leistungen benotet wurden, welche die Zeugnisnote positiv beeinflussten beziehungsweise stabilisierten. Aus dem Interviewzitat von Lehrkraft 9 geht jedoch die grundsätzliche Problematik hervor, die mit einer Benotung verbunden sind. Für eine Zuweisung der Testleistung zu einer Schulnote ist ein Maßstab erforderlich, den die Lehrkräfte selbst festlegen müssten. Die Lernstandserhebung sieht aufgrund ihrer testtheoretischen Konzeption einen solchen Bewertungsmaßstab nicht vor. Auf diese Weise wird der Notenschlüssel willkürlich bestimmt. Indem ein Maßstab verwendet wird, der den üblichen Klassendurchschnitt wiederspiegelt, wie Lehrkraft 9 berichtete, werden die Ergebnisse zudem verfälscht. Die Aussagekraft für Lehrperson und Schüler reduziert sich immens und beschränkt sich auf den Notenwert. Dies stellt eine selektive Auswertungsform dar und das eigentliche diagnostische Ziel der Lernstandserhebung, die kriteriale Diagnose der Schülerkompetenzen, gerät in den Hintergrund.

Anregung durch Aufgabenformate

Im Kontext der Innovationsfunktion sollen die Lernstandserhebungen die Lehrkräfte mit neuartigen Aufgabenformaten und -formulierungen konfrontieren. Inwiefern sie daraus Anregungen zur Weiterentwicklung ihres eigenen Unterrichts erhielten, wird anhand der folgenden Beispiele ersichtlich:

> *LK 2: Was ich Mittwoch vorschlagen möchte, ist, dass man vielleicht die eine oder andere Aufgabe aus diesem Jahr oder aus den früheren Jahren vielleicht einmal als Vergleichsarbeit nutzt, weil sie bestimmte Richtungen anstößt, die wir bislang noch nicht so gesehen haben. […] Wir könnten uns durchaus anregen lassen durch so etwas, fände ich gut […]. (LK 2, 00:24:49-9)*
>
> *SL 8: Das Aufgabenmaterial der [Lernstandserhebung] kann dabei als Anregung bei der Erstellung von Unterrichtsmaterial dienen. (SL 8, Nachgespräch per E-Mail)*
>
> *LK 4: Davon bin ich überzeugt, dass Lehrkräfte, und das werde ich auch tun, mich an diesen Aufgabenformaten auch orientiere und auch deren Vorteile erkenne für den Unterricht. (LK 4, 00:13:37-0)*

Aus den Äußerungen der zitierten Lehrkräfte gehen zwei verschiedene Sichtweisen in Bezug auf Anregungen durch die Aufgabenformate hervor. Die Lehrkraft 2 schlug vor, die Aufgaben bei der Konzeption der Vergleichsarbeiten zu berücksichtigen. Sie bezog sich hierbei insbesondere auf die Aufgabenformate. Die Anregung war somit eher konzeptioneller als

inhaltlicher Art. Demgegenüber war bei der Lehrkraft 4 und dem Schulleitungsmitglied 8 eine Anregung mit Blick auf kompetenzorientierte Aufgabenstellungen zu erkennen, die den Unterricht bereichern soll. Hiermit ist die inhaltliche Ausrichtung der Items impliziert und nicht die Übernahme von beispielsweise Multiple-Choice-Aufgaben.

Insgesamt bestätigten acht befragte Personen, dass sie die Aufgaben für ihren eigenen Unterricht anregend fänden. Allerdings hatten lediglich zwei Befragte ihre Anregungen zum Erhebungszeitpunkt umgesetzt. Deren Maßnahmen gehen aus folgenden Äußerungen hervor:

> *SL 1: Aber ich habe nach dem letzten Lernkompetenzdurchgang unserer Schule gedacht, in diesem Jahr bereite ich die Vergleichsarbeit für die Klasse 6 vor. Und dann hat das auch auf die Aufgabenstruktur eingewirkt. Ich habe versucht, in dieser Vergleichsarbeit auch aus diesen technischen Fertigkeiten herauszugehen und mehr die anderen Kompetenzen mit in das Spiel zu bringen. (SL 1, 00:14:12-5)*

> *SL 9: Allerdings, gerade bei jüngeren Kollegen und Kolleginnen, macht sich mit diesen ganzen Lernstandserhebungen so ein bisschen die Stimmung breit, "Learning on the Test". Das heißt, Vorbereitung auf den Test, Formatkenntnisse etc. werden wichtiger, damit man eben gut abschneidet. Und das sind Entwicklungen, die finde ich nicht gut! (SL 9, 00:06:10-4)*

Demzufolge griff Schulleitungsmitglied 1 ebenfalls die Idee auf, die schulinterne Vergleichsarbeit an den Aufgaben der Lernstandserhebung zu orientieren. Allerdings bezog er sich nicht auf die Aufgabenformate, sondern auf eine verstärkte kompetenzorientierte Ausrichtung der Parallelarbeit. Der Schulleitungsvertreter 9 berichtete demgegenüber von einem verstärkten Einsatz der Aufgabentypen bei aufgetretenen Defiziten in den Testresultaten. Er setzte dies allerdings nicht selbst um, sondern beobachtete diese Entwicklung bei Kollegen.

Den beiden Ausführungen können zwei verschiedene Zielsetzungen zugeordnet werden. Die Maßnahmen dienen dazu, Leistungsschwächen zu reduzieren. Allerdings ist zu unterscheiden, wie diejenige Lehrkraft diese Schwächen interpretiert. Entweder müssen die Schüler inhaltlich in Bezug auf einen Kompetenzbereich stärker gefördert werden oder der Umgang mit den Aufgabenformaten muss verstärkt eingeübt werden. Dementsprechend handelt es sich bei der Aussage von Schulleitung 9 um letzteren Fall, indem im Sinne eines Teaching to the Test die Schüler mit den Testformaten konfrontiert wurden, um mit den Aufgabenstellungen effektiv umgehen zu können.

Da ausschließlich einer von neunzehn Befragten, die den Test durchführten, die Anregungen in seinem Unterricht selbst umgesetzt hat, sollen an dieser Stelle die angeführten Gründe für die geringe Einbindung der Aufgaben in die Unterrichtskonzeption dargelegt werden.

LK 3: Aber das sind auch Aufgaben, die würde ich so nie stellen. Also ich würde nie Aufgaben stellen, wo die nur Richtig oder Falsch anzukreuzen haben. (LK 3, 00:19:11-3)

LK 6: Leider gibt es für den Deutschunterricht wenig bis gar keine Materialien. Das selbständige Erstellen dieser Materialien ist sehr zeitaufwändig, daher kaum möglich. (LK 6, Nachgespräch per E-Mail)

LK 7: Sondern man ist ja schon an das Buch gebunden als Vorlage. […] Das heißt also, ich habe den Unterricht jetzt nicht umstellen können […] für die Schüler, denen das natürlich gelegen hat, die Art des Abfragens. Weil einfach da die Materialien nicht sind. (LK 7, 00:15:31-8)

Als ein zentrales Hindernis wurde das Aufgabenformat der Multiple-Choice-Items angeführt, deren Einsatz im Unterricht nicht sinnvoll sei. Dieses Argument berücksichtigt nicht, dass die Lernstandserhebung nur zu einem gewissen Teil aus geschlossenen Aufgaben besteht und durchaus auch offene Items beinhaltet. Zudem könnten die geschlossenen Items in offene Lernaufgaben transformiert werden, was jedoch einen erhöhten Arbeitsaufwand für die Lehrkraft impliziert. Dieser sei laut den angeführten Beispielaussagen zu hoch und einsetzbares Material sei noch nicht zur Verfügung gestellt worden. Zudem trat bei der Lehrkraft 7 die Sichtweise signifikant hervor, den Unterricht eng an den vorgegebenen Lehrmaterialien orientieren zu müssen, so dass zeitlich wie auch didaktisch kein großer Spielraum für andere Aufgabenformulierungen bleibe.

Individuelle Förderung

Auf Grundlage einer Reflexion der Schülerleistungen lassen sich Maßnahmen zur Förderung von einzelnen Schülern oder Kleingruppen ableiten. Sechs Befragte erkannten ein förderdiagnostisches Potenzial. Jedoch konnte lediglich bei einem Befragten die tatsächliche Umsetzung einer Maßnahme beobachtet werden.

LK 5: Also ich weiß jetzt, dass bestimmte produktive Aufgaben ohne Probleme von bestimmten Schülern ohne Hilfestellung bewältigt werden, während andere dann vielleicht doch eher in kleineren Gruppen zusammenarbeiten sollten […]. Und das habe ich jetzt auch gemacht. Also ich hab eine Referendarin noch mit drin und da haben wir auch ein bisschen daran gearbeitet, also auch Aufgabenformulierungen dann stärker zu differenzieren. (LK 5, 00:07:08-7)

Die Aufgabenstellungen im Unterricht wurden somit verstärkt binnendifferenzierend konzipiert, so dass die Schüler leistungshomogenen Kleingruppen zugeordnet wurden und entsprechend mehr Hilfestellung erhielten beziehungsweise anspruchsvollere Aufgaben bearbeiteten.

Bedeutsam ist im Bereich der Förderung die Frage, warum die Lehrkräfte zwar deren Potenzial erkannten, aber keine Maßnahmen dahingehend geleitet haben. Exemplarisch geben folgende Interviewausschnitte zu dieser Problematik Aufschluss:

LK 2: […] [Ich] habe ja nur einen summarischen Bericht über die gesamte Klasse zurückbekommen, ich müsste jetzt Schülerprofile einzeln nochmal erstellen, um zu sagen: "Da und da liegen deren Defizite." Das ist eine Zeitfrage! (LK 2, 00:12:07-9)

SL 5: Also, da tue ich mich momentan noch ein bisschen schwer, weil ich jetzt-, da müsste ich jetzt bei dem Einzelnen immer genau wissen, der und der kann das und das gar nicht und deshalb-. Also wenn man sich ganz viel Mühe macht, könnte man da schauen, wie kann ich gezielt den Einzelnen fördern? Das geht aber nur, finde ich, in kleinen Lerngruppen. (SL 5, 00:07:01-3)

SL 4: Das wissen wir, dass das ein Problem ist, und da wissen wir auch, klar, wir könnten mehr binnendifferenzierend arbeiten. […] Das wissen wir und das ist auch ein Ziel, an dem wir weiter arbeiten müssen. Das hat auch die Schulinspektion ergeben und so. (SL 4, 00:32:27-7)

Indem die Rückmeldungen lediglich Durchschnittswerte der Gesamtklasse sowie Verteilungen der Leistungsprofile innerhalb der Lerngruppe wiederspiegeln, obliegt es der Lehrkraft für die Schüler den individuellen Kompetenzstand zu diagnostizieren und Fördermaßnahmen abzuleiten. Diese Aufgabe setze nach Ansicht der Befragten einen immensen Arbeitsaufwand voraus, der aufgrund der Klassenstärke nicht zu leisten sei.

Der Schulleitungsvertreter 4 führte indes an, dass die Problematik einer bislang mangelhaften differenzierenden Ausrichtung des Unterrichts in der Schule bekannt sei und durch die Lernstandserhebung erneut bestätigt worden sei. Dies habe jedoch keine Initiative zur Handlungsbereitschaft ausgelöst. Der Test wirkte keineswegs als Katalysator. Der Befragte begründete dies mit der Langfristigkeit des Prozesses, der eine Professionalitätsentwicklung im Kollegium bedinge.

Anregungen durch Testinhalte

Dem letzten Bereich „Anregung durch Testinhalte“ werden Äußerungen der Probanden zugeordnet, aus denen eine Weiterentwicklung der zu unterrichtenden Inhalte und Kompetenzen hervorgeht. Die zugehörigen Handlungskonzepte sind hierbei vielfältig und können den folgenden Interviewausschnitten entnommen werden:

LK 9: [In] einem Fall das eine Thema habe ich dann wirklich […] nach der Arbeit im Prinzip nochmal aufgegriffen und nochmal bearbeitet, weil es halt scheinbar definitiv zu kurz kam. (LK 9, 00:17:26-3)

LK 6: Aber eben diese Hörverstehensübungen, ich denke, die sind allgemein im Fach Deutsch relativ neu, dass die da auch Anklang finden beziehungsweise in den Unterricht auch einziehen sollten. (LK 6, 00:07:20-0)

LK 1: Und der Unterschied zu A1, A2 ist vor allem eben auch, dass ich komplexere Sätze bilden kann, dass ich sentence connectors, linking words benutze […].Und das ist auch in unserem Lehrwerk so und das hatten wir dann durchaus auch gemacht schon. Aber ich habe das jetzt nochmal, denke ich, mehr betont. (LK 1, 00:04:58-1)

SL 7: Und insofern habe ich natürlich als Konsequenz erst mal gezogen, dass man unter Umständen Schüler auch auf diesen Zeitdruck vorbereiten muss. […] Insofern, dass man schon in Arbeitsphasen sehr viel stärker auf Einhaltung von Zeitvorgaben achtet und den Schülern auch klar macht: "Ihr müsst mit der Zeitvorgabe einfach klar kommen. Ihr müsst da etwas zum Abschluss bringen." (SL 7, 00:05:27-9)

Aus der Aussage der Lehrkraft 9 geht hervor, dass insbesondere die Reflexion defizitärer Ergebnisse zu Anregungen geführt haben, welche Themen- und Kompetenzbereiche nochmals vertiefend im Unterricht betrachtet werden sollten. In diesem Beispiel wurde eine Wiederholung in der getesteten Lerngruppe durchgeführt. Andere Befragte berichteten von der Absicht, zukünftig einen erhöhten Schwerpunkt auf die Kompetenzbereiche zu legen. Von insgesamt vier Probanden, die eine solche Anregung schilderten, setzte sie lediglich Lehrkraft 9 praktisch um. Dennoch kann in den anderen Fällen eine Unterrichtsentwicklung stattgefunden haben, da den Lehrkräften zumindest bewusst wurde, welche Kompetenzbereiche stärker gefördert werden sollten. Es ist durchaus möglich, dass diese Anregungen außerhalb des Erhebungszeitraumes längerfristig umgesetzt wurden.

Interessant ist zudem die Schilderung von der Lehrkraft 6. Sie war durch die Lernstandserhebung auf Kompetenzbereiche aufmerksam geworden, denen sie bislang in ihrer Unterrichtskonzeption noch keine größere Beachtung beigemessen hatte. Aus der großen Gewichtung dieser Bereiche in der Lernstandserhebung folgerte sie, dass diese Kompetenzfelder im Kontext der Bildungsstandards eine stärkere Bedeutung einnehmen sollten. Insgesamt berichteten vier Befragte von Anregungen dieser Art; neben den genannten Bereichen des Hörverstehens wurden auch die szenische Umschreibung einer Textsituation und die Interpretation von Diagrammen angeführt. Jedoch setzte lediglich ein Befragter eine entsprechende Maßnahme um, indem er in der Unterrichtseinheit „Lyrik“ vertonte Gedichte thematisierte und somit den Bereich des Hörverstehens intensiver fokussierte.

Die Lehrkraft 1 ist hingegen auf die Notwendigkeit einer stärkeren Differenzierung zwischen den einzelnen Anforderungsstufen im Fremdsprachenunterricht aufmerksam geworden. Zwar war dies für sie kein neuartiges Konzept, doch die Lernstandserhebung diente als Verstärkung, auf diese Differenzierung gezielter zu achten.

Die im Interviewzitat beschriebene Maßnahme von dem Schulleitungsvertreter 7 zielte indes weniger auf fachliche Kompetenzbereiche ab, sondern auf überfachliche Fähigkeiten im Kontext der Arbeits- und Selbstkompetenz. Da die Schüler mit den Zeitregularien des Tests nicht zurechtkamen, wurde der Befragte darauf aufmerksam, zukünftig stärker auf Zeitvorgaben und deren Einhaltung in alltäglichen Unterrichtssituationen zu achten und setzte dies dementsprechend auch um.

Insgesamt konnten bei neun Probanden Anregungen inhaltlicher Art festgestellt werden. Tatsächlich realisierte Maßnahmen wurden hingegen nur bei vier Befragten beobachtet. Hierfür konnten aus den Interviews drei wesentliche Gründe ermittelt werden.

LK 4: Ich würde sagen, diese Dinge der Gewinnung von Ergebnissen für mich oder von Folgerungen, die werde ich jetzt erst in den nächsten Wochen nachvollziehen. Die haben jetzt den aktuellen Unterrichtsablauf noch nicht beeinflusst, aber sicherlich […] die Planung und die Herangehensweise bei neuen Klassen im Vergleich. (LK 4, 00:10:00-5)

SL 8: Nein, wie gesagt, der Vorsatz bestand, aber es ist nicht dazu gekommen. Aber man hätte natürlich durchaus die Möglichkeit, wenn man sich intensiv damit beschäftigt, dann auch für sich und für seinen Unterricht die Konsequenz zu ziehen. Die Möglichkeit besteht. Aber wie gesagt, Alltag. Schulalltag lässt einem häufig nicht die Zeit. (SL 8, 00:06:43-4)

SL 11: Also sobald Sie über dem Landesdurchschnitt sind, sind Sie erst mal zufrieden mit sich und der Welt. Da gibt es ja überhaupt keinen Handlungsbedarf, muss man mal so sagen. (LK 11 + SL 11, 00:25:36-8)

Die Äußerung von Lehrkraft 4 bestätigt die Möglichkeit die Lernstandserhebung über einen langfristigen Zeitraum zu nutzen und konkrete Maßnahmen nicht sofort, sondern beispielsweise erst in einer anderen Lerngruppe derselben Klassenstufe zu ergreifen. Andererseits erfordert dies eine längerfristige Verankerung der Anregungen im Bewusstsein, damit nach einem größeren Zeitraum erneut darauf zurückgegriffen werden kann. Dies hängt insbesondere von der Intensität der Rezeption und Reflexion der Testergebnisse ab.

Als einen weiteren Grund für eine nicht existente Handlungsbereitschaft wurde die fehlende Zeit im Schullalltag von mehreren Befragten angeführt. Letztlich ist mit einer intensiven Auswertung der Tests, der Ableitung und der Umsetzung zugehöriger Maßnahmen ein enormer Zeitaufwand verbunden, den viele Lehrkräfte nicht bereit waren zu übernehmen.

Als häufigstes Hindernis wurde jedoch von sechs Probanden die Zufriedenheit mit den Schülerleistungen angeführt. Indem die Ergebnisse positiv wahrgenommen wurden, ergab sich für die jeweilige Lehrkraft keine Notwendigkeit zu Veränderungen im eigenen Unterricht. Vielmehr wurde die Rückmeldung als Bestätigung des eigenen Handelns verstanden, welche keine weiteren Maßnahmen erfordert. Diese Haltung wurde durch die Verortung der Klassenergebnisse am korrigierten Landesmittelwert verstärkt. Sobald die Schülerleistungen dem landesweiten Durchschnittswert entsprachen oder ihn übertrafen, bestand die Tendenz, die Nutzung auf diese Informationen zu beschränken und keine Handlungsschritte einzuleiten.

Bislang wurden die beobachteten Anregungen und die eventuelle Umsetzung von Maßnahmen nach einzelnen Bereichen sortiert vorgestellt. Im Folgenden sollen die Ergebnisse dahingehend gemeinsam analysiert werden, inwiefern sie eine tatsächliche Unterrichtsentwicklung befördert haben. Hierzu es ist sinnvoll, die Intensität der Anregungen und Äußerungen dimensional zu erfassen und die Probanden erneut in vier Intensitätskategorien einzuteilen (vgl. Tabelle 13).

Kategorien zur Erfassung der Aktionsintensität für die Unterrichtsentwicklung		
Kategorie	**Beschreibung**	**Zugeordnete Personen**
keine Aktion	Es sind keine Anregungen oder Aktionen zur Unterrichtsentwicklung festzustellen.	LK 3, LK 7, LK 8, LK 11, LK 12, SL 9
geringe Aktion	Die Lehrkraft hat verschiedene Anregungen zur Unterrichtsentwicklung durch Aufgabenformate, -inhalte sowie im Bereich der individuellen Förderung gewonnen, diese aber (noch) nicht umgesetzt.	LK 1, LK 4, LK 6, SL 3, SL 4, SL 8, SL 12
mittlere Aktion	Die Lehrkraft hat eine Maßnahme zur Unterrichtsentwicklung ergriffen.	LK 2, LK 5, LK 9, LK 10, SL 1, SL 5,
intensive Aktion	Die Lehrkraft hat mindestens zwei verschiedene Maßnahmen zur Unterrichtsentwicklung ergriffen.	-

Tabelle 13: Kategorien zur Erfassung der Aktionsintensität für die Unterrichtsentwicklung

Es erfolgte eine Einteilung in keine, geringe, mittlere und intensive Aktion. Wenn die Probanden keine Anregungen für die Weiterentwicklung ihres eigenen Unterrichts aus den Ergebnissen der Lernstandserhebungen ziehen konnten, wurden sie der Kategorie „keine Aktion“ zugeordnet. Hierzu zählen auch die Lehrkräfte, welche lediglich die Schülerleistungen in den Tests bewertet haben. Dies stellt lediglich eine summative Leistungsmessung in Form einer Notenvergabe dar und hat keine formative Unterrichtsentwicklung zur Folge, so dass diese Maßnahme die Unterrichtsqualität nicht verbessert.

Bei der geringen Aktion haben die Befragten lediglich Anregungen in den einzelnen Bereichen gewonnen. Eine Umsetzung ist jedoch nicht erfolgt, so dass zu diesem Zeitpunkt keine Unterrichtsentwicklung in der betreffenden Klasse einsetzte. Dennoch werden Anregungen als eine geringe Aktion verstanden, da sie das professionelle Verständnis erweitern und somit zu einer indirekten Unterrichtsentwicklung führen können. Es ist durchaus möglich, dass diese Probanden zu einem späteren Zeitpunkt und in anderen Lerngruppen entsprechende Handlungen einleiten, die in dieser Untersuchung nicht berücksichtigt werden konnten.

Bei der mittleren Aktion wurde eine gewonnene Anregung tatsächlich umgesetzt, so dass von einer direkten Unterrichtsentwicklung ausgegangen werden kann. Zu dieser Kategorie wurde ebenfalls eine intensive inhaltliche Auswertung der Lernstandserhebungen mit der Schülergruppe gezählt, denn hierbei werden einzelne Kompetenzbereiche nochmals thematisiert und gefördert.

Erfolgte eine quantitativ höhere Umsetzung von Anregungen im Unterricht, wurde eine Zuordnung zur intensiven Aktion vorgenommen.

Erneut muss darauf hingewiesen werden, dass die Beschreibungen der Kategorien lediglich die Spannbreite des über die Interviews ermittelten Aktionsverhaltens wiederspiegeln und nicht als allgemeingültig interpretiert werden können. Zudem wurden nur diejenigen Probanden den Kategorien zugeordnet, die selbst an einer Lernstandserhebung teilgenommen haben.

Insgesamt haben etwa zwei Drittel der Befragten, welche selbst an einem Test teilnahmen, keine Änderungen in ihrem Unterricht vorgenommen. Davon gewannen insgesamt sechs Personen keine Anregungen für eine Unterrichtsentwicklung aus den Ergebnissen. Daraus lässt sich folgern, dass mit den Lernstandserhebungen und den Rückmeldungen oftmals eine Weiterentwicklung des Unterrichts nicht befördert wurde. Die Lehrkräfte, die tatsächlich Handlungen in ihrem Unterricht vorgenommen haben, ergriffen lediglich eine Maßnahme. Zudem bezogen sich diese Handlungen nicht auf verschiedene Bereiche der Unterrichtsentwicklung und sind eher punktuell zu deuten.

Aus der Verteilung der Probanden lässt sich weiterhin eine tendenzielle Vermutung ableiten, dass Schulleitungsmitglieder eher Anregungen ableiten. Des Weiteren wird aus der Zuordnung der Befragten zu den Kategorien bezüglich der Reflexion und Aktion deutlich, dass eine intensive Reflexion keineswegs eine mittlere oder intensive Aktion zur Folge hat. Ein konstanter Nutzungsprozess war bei diesen beiden Phasen lediglich bei fünf Probanden (davon bei 4 Schulleitungsvertretern) festzustellen. Bei zwei Befragten war eine Erhöhung der Nutzungsintensität von einer geringen Reflexion zu einer mittleren Aktion zu konstatieren. Prägnanter ist der Abfall der Nutzungsstärke bei insgesamt zwölf Lehrkräften und Schulleitungsvertretern. Bei fünf Personen konnte sogar eine Reduzierung um mindestens zwei Intensitätsstufen konnotiert werden. Folglich scheint der Übergang von der Phase der Reflexion zur Aktion für die teilnehmenden Lehrkräfte eine Herausforderung dargestellt zu haben, die durch verschiedene Faktoren primär negativ beeinflusst wurde. Die Handlungsbereitschaft führte in nur sechs Fällen zu einer tatsächlichen Umsetzung. Die Gründe für die Verhinderung einer Umsetzung von Maßnahmen wurden in den vorangegangenen Ausführungen zu den Ergebnissen der Aktion ausführlich dargelegt.

9.3.2 Organisationsentwicklung

Neben dem Anstoß von Unterrichtsentwicklung kann der Nutzungsprozess auch die Organisationsentwicklung befördern. Dies betrifft speziell die organisatorische Abwicklung der Lernstandserhebung in der Schule oder hat Auswirkungen auf bereits vorhandene Organisationsstrukturen beziehungsweise auf die Initiierung neuer Strukturen. Zum ersten Bereich lassen sich aus den Interviewaussagen folgende Erkenntnisse ableiten.

SL 2: Dann hat [der Koordinator] auch die komplette Organisation übernommen. Das werden wir im nächsten Jahr anders machen. Wir haben jetzt pro Fach Beauftragte, die das organisieren helfen, damit es nicht zu viel für einen wird. (SL 2, 00:04:47-9)

SL 5: Also wir hatten das am Anfang, weil es immer so ein bisschen schwierig war, wer organisiert das Ganze? Und das hat unser stellvertretender Schulleiter am Anfang gemacht, auch diese Eingabe. […] [Und] jetzt möchten wir aber in Zukunft, dass das eine konkrete Person macht. Das heißt, das hat dann aus der Schulleitung da niemand mit direkt etwas zu tun. (SL 5, 00:16:45-7)

LK 1: Im letzten Jahr haben die Kollegen das selber korrigiert und haben gesagt: "Gut, wir probieren das aus." Und haben dann festgestellt, dass es eben doch eine erhebliche Mehrarbeit ist. Und dieses Mal war es halt so, dass versprochen wurde, dass die Korrekturen nach außen gehen, also dass nicht wir das machen, sondern andere Personen. (LK 1, 00:00:48-7)

LK 9: Wir haben es jetzt dieses Jahr und auch letztes Jahr so gemacht, wir haben sie im Prinzip auf das ganze Kollegium oder auf die ganze Fachschaft die Korrektur aufgeteilt. Also auch die Kollegen, die nicht in den Klassen waren, haben auch korrigiert. Und dann hatte jeder so sieben, acht solche Lernstandserhebungen. (LK 9, 00:08:49-0)

Aus den Beispielen wird ersichtlich, dass es sich um geplante oder teilweise bereits realisierte Maßnahmen zur Verbesserung der organisatorischen Abwicklung der Lernstandserhebungen handelte. Während in der Schule 2 zukünftig jeweils ein Koordinator pro Fach bestimmt wird, der neben den organisatorischen Fragen auch als inhaltlicher Ansprechpartner für die Tests zur Verfügung steht, bemühte sich die Schulleitung 5, sich mit der Betreuung der Lernstandserhebung zu entlasten und die zugehörigen Aufgaben einer anderen Person zuzuweisen.

Die Lehrkräfte 1 und 9 thematisierten weiterhin Änderungen in der Korrekturpraxis aufgrund des als zu hoch empfundenen Arbeitsaufwands. Die beiden Schulen schlugen hierbei verschiedene Wege ein. In der Schule 1 wurden die Lehrpersonen gänzlich von der Korrektur entbunden, indem externe Personen wie Studenten die Korrektur übernahmen. Dies hatte jedoch den Nachteil, dass sich die betreffenden Lehrpersonen nicht intensiv mit den Testinhalten auseinandersetzten, was eine Reflexion der Ergebnisse massiv erschwerte. In der Schule 9 wurde hingegen die Korrektur auf das gesamte Kollegium eines Fachs aufgeteilt. Dies bedeutete einerseits unabhängig von der eigenen Teilnahme einen Mehraufwand

für alle Lehrpersonen des Faches. Andererseits wurden alle Lehrkräfte mit dem Testkonzept und -inhalten konfrontiert, so dass sie eventuell bei einer späteren eigenen Teilnahme vertrauter mit dem Messinstrument umgehen können.

Letztlich sind alle genannten Maßnahmen aus den Erfahrungen im Umgang mit den Lernstandserhebungen entstanden, verbunden mit dem Wunsch, die organisatorische Abwicklung effizienter zu gestalten.

Demgegenüber wurden Handlungen zur Verbesserung bereits etablierter Organisationsstrukturen oder die Initiierung neuer Strukturen nur selten angeführt. Lediglich drei Befragte äußerten die Wahrnehmung eines möglichen Potenzials für eine Verbesserung der Kooperationsstrukturen, wie in den folgenden Zitaten ersichtlich ist:

> *SL 2: Wenn auf der einen Stelle die persönliche Zusammenarbeit gut läuft, ist man auch viel eher bereit, über dieses Vehikel Lernstandstest […] diese Teamfähigkeit bei einer anderen Gelegenheit […] auszubauen. Ich glaube insofern, zusammenfassend zu sagen, dass die Lernstandstests ein Instrument sind, mit dem Teamfähigkeit oder Teamarbeit gefördert werden kann. (SL 2, 00:28:34-1)*

> *SL 3: Sie spielen momentan in den Fachkonferenzen noch keine Rolle. Genau diesen Punkt haben wir gerade bei der letzten Fachkonferenz angesprochen, dass wir gesagt haben: "Ok, jetzt haben wir mal geguckt, wie das so aussieht alles und welche Auswirkungen hat es denn auf unseren Unterricht?" Und da wollen wir jetzt, das steht sozusagen auf der Agenda für das nächste Schuljahr da ranzugehen […] und da den Lernstandserhebungen einen größeren Wert beizumessen als bisher. (SL 3, 00:04:02-4)*

Das Schulleitungsmitglied 2 führte demzufolge die Hoffnung auf eine Festigung und Vertiefung bereits vorhandener Teamstrukturen mithilfe der Arbeit an den Lernstandserhebungen an. Dass die Lernstandserhebung jedoch auch dazu geeignet ist, neue Kooperationsstrukturen einzurichten, konnte an keiner der untersuchten Schule festgestellt werden. Im zweiten Beispiel wurde hingegen die Absicht ersichtlich, zukünftig den Lernstandserhebungen eine größere Bedeutung im Schulalltag einzuräumen. Dies würde eine Verknüpfung mit bereits vorhandenen Organisationsstrukturen erfordern. Jedoch stellte dieser Gedanke lediglich eine zukünftige Vision dar, so dass keine Ideen zur Umsetzung im Befragungszeitraum vorlagen.

Als Gründe für die geringfügige Ableitung von Maßnahmen aus den Lernstandserhebungen für die Weiterentwicklung von Organisationsstrukturen wurde von den Probanden wiederholt der geringe Stellenwert der Tests im schulischen Alltag, die nicht flächendeckende Teilnahme aller Fachlehrkräfte an den Tests sowie das individuelle Bedürfnis, zusätzlichen Arbeitsaufwand so gering wie möglich zu halten, angeführt. Diese Gründe stehen im Einklang zu den bereits genannten Faktoren, welche die Rezeption und Reflexion negativ beeinflusst haben.

9.3.3 Personalentwicklung

Ergriffene Maßnahmen als Konsequenzen zu den Lernstandserhebungen im Bereich der Personalentwicklung betreffen insbesondere die Weiterentwicklung der fachlichen und diagnostischen Professionalität der jeweiligen Lehrkraft. Dies stellt entweder einen unbewussten Entwicklungsprozess dar oder wird von der betreffenden Person initiiert. Aus den Interviews mit den schulischen Akteuren konnten Hinweise auf eine forcierte Personalentwicklung herausgefiltert werden.

LK 4: Ich denke, dass wir in dem Bereich der Diagnostik auf jeden Fall einen großen Bedarf haben. Ich persönlich für mich sehe das auf jeden Fall so und ich nehme das auch so wahr, dass es im Kollegium so gesehen wird. […] Von daher ist dort in dem Fall Fortbildungsbedarf vorhanden. (LK 4, 00:34:11-4)

LK 8: Und das finde ich auch ein Problem, wenn wir jetzt die Bildungsstandards kriegen, müssen wir ja andere Formen von Arbeiten oder von Abtestmöglichkeiten machen und da denke ich, ist einfach Fortbildungsbedarf an den Schulen. (LK 8, 00:11:27-1)

SL 11: Ach ja, ich würde mich mit den Aufgabenformaten nicht befassen in dem Sinne, wenn sie mir das nicht in einer Lernstandserhebung Klasse 8 Englisch reinpacken würden. (SL 11 + LK 11, 00:29:53-8)

SL 12: Das ist eine Aufgabe, wo verschiedenste Bereiche angesprochen werden […]. Und ich denke, für mich war das schon ein Signal, dass wir da im Mathematikunterricht verstärkt darauf achten müssen. (SL 12 + LK 12, 00:25:07-0)

Wie aus den Äußerungen von den Lehrkräften 4 und 8 ersichtlich wird, wurde durch die Auseinandersetzung mit den Lernstandserhebungen ein Fortbildungsbedarf bei sich oder dem Kollegium erkannt. In beiden Fällen ging die thematische Ausrichtung einer gewünschten Fortbildung über das Kerngebiet des Tests hinaus und betraf allgemeine Möglichkeiten eines an den Bildungsstandards ausgerichteten Diagnostizierens und Bewertens im Unterricht.

Insgesamt äußerten fünf Lehrpersonen einen Fortbildungsbedarf, wobei sich bei zwei Befragten das Bedürfnis auf eine Unterweisung speziell zur Nutzung der Ergebnisse aus den Lernstandserhebungen bezog. Hierbei wurde die Unzufriedenheit mit den Erkenntnissen der eigenen Auswertung deutlich, so dass eine effektivere Nutzung gewünscht wurde. In allen Fällen wurde jedoch lediglich ein Fortbildungsbedarf genannt, aber zugehörige konkrete Maßnahmen zur Umsetzung wurden innerhalb des Erhebungszeitraums nicht ergriffen.

Bei den Schulleitungsmitgliedern 11 und 12 wurde hingegen eine direkte Weiterentwicklung der Professionalität deutlich, da sich beide Personen im Zuge der Testnutzung verstärkt mit neuen Aufgabenformaten und -inhalten auseinandergesetzt und daraus Anregungen gewonnen haben. Auf diese Weise wurden sie mit Aufgabenformen eines kompe-

tenzorientierten Unterrichts konfrontiert, was ihre bisherigen Professionskompetenzen gestärkt und erweitert hat. Hierzu zählen insbesondere alle inhaltlichen Anregungen, welche die Lehrpersonen aus der Reflexion erhalten haben. Die einzelnen Dimensionen solcher Anregungen bei den Befragten wurden in Abschnitt 9.3.1 beschrieben.

9.3.4 Profilierung der Schule

Bislang wurden die Maßnahmen, welche nach der Auswertung der Testergebnisse ergriffen wurden, den Schulentwicklungsbereichen Unterrichts-, Organisations- und Personalentwicklung zugeordnet. Dies ermöglichte eine Strukturierung der Ergebnisse für die Nutzungsphase der Aktion.

Einige Schulen entwickelten zusätzlich Konzepte zur Verbesserung der Wahrnehmung der Schulqualität. Diese Handlungen lassen sich nicht in die oben genannten Schulentwicklungsbereiche einordnen, da sie keine direkte Weiterentwicklung zur Folge hatten, sondern vielmehr eine Rechtfertigungsfunktion verfolgten. Ihnen lag die Überzeugung zugrunde, dass sich aus den Testergebnissen der teilnehmenden Klassen ein Leistungsbild über die Qualität der jeweiligen Schule ableiten lasse:

> *SL 3: […] [Also] man führt diese Lernstandserhebung durch und die fallen alle positiv aus, dann würde ich behaupten […], ist es ein Indiz dafür, dass in der Schule ein nachhaltiges Lernen stattfindet. (SL 3, 00:14:53-2)*

Die Erkenntnisse aus der Reflexion wurden vom Mikrokosmos des Unterrichts auf die Arbeit der gesamten Schule verallgemeinernd übertragen. Dies wurde insbesondere von den Schulleitungen vorgenommen. Einerseits können solche Erkenntnisse weiterführend zu einer Schulentwicklung beitragen, indem Maßnahmen zur Behebung von Defiziten oder zur Bestärkung von positiven Leistungen auf Schulebene ergriffen werden. Andererseits kann eine Zufriedenheit mit den Ergebnissen auch zu dem Bedürfnis führen, die Schule damit zu profilieren und die Wahrnehmung sowie das Image der Schule im äußeren Umfeld positiv zu beeinflussen.

> *SL 6: Und jetzt kriegen wir diese Rückmeldung und wenn man darauf verweisen kann und auch bei Schulvorstellungen oder bei sonstigen Konferenzen darauf verweisen kann oder auch mal gegenüber der Presse, dann ist das eben auch gut für das Image der Schule. Und hat dann auch einen Werbeeffekt für uns. Dann müssen wir uns vielleicht nicht mehr so sehr kümmern darum, dass wir zu wenig Schüler bekommen. (SL 6, 00:12:00-0)*

Mit der Veröffentlichung der Testergebnisse könne nach dieser Meinung die Wahrnehmung der Schulqualität massiv verbessert werden. Dies wirke sich nicht nur auf das interne Kollegium und die Schülerschaft im Kontext eines eventuell verbesserten Arbeitsklimas aus, son-

dern solle insbesondere potentielle Schüler sowie deren Eltern erreichen und die Schulentscheidung beeinflussen. Auf Nachfrage teilte das Schulleitungsmitglied 6 später mit, dass in der regionalen Zeitung ein Artikel über das positive Abschneiden der Schule an den Lernstandserhebungen erschienen ist. Alle anderen untersuchten Schulen haben auf eine Weitergabe der Testergebnisse mit dem Ziel einer Positionierung im direkten Vergleich zu Konkurrenzschulen verzichtet und sind damit der offiziellen Bitte des Landesinstituts gefolgt, ein inoffizielles Ranking zu verhindern. Als Gründe für die Ablehnung einer solchen Veröffentlichung wurden von den Befragten die Unklarheit des Bewertungsmaßstabes „einer guten Leistung“ sowie die geringe Teilnahme der Lerngruppen prozentual zur Gesamtschülerzahl einer Schule angeführt.

9.4 Evaluation

Der Nutzungsprozess nach dem Modell von Helmke (vgl. Helmke A. , 2004, S. 100) schließt letztlich mit einer Evaluation der ergriffenen Maßnahmen ab. Dies kann eine formalisierte Evaluation sein oder auch eine Reflexion durch die Schüler beziehungsweise die Lehrperson selbst. Es soll der Erfolg der Handlungen bewertet werden, so dass im Idealfall aus dieser Erkenntnis weitere Schritte zur Qualitätssicherung und -entwicklung formuliert werden können.

Ein Befragter, welcher im Anschluss an die Phase der Aktion eine Evaluation vornahm, ist der Schulleitungsvertreter 12, der selbst mit einer Klasse an der Lernstandserhebung teilgenommen hatte. Wie in Abschnitt 9.2.2 im Kontext der Attribuierung der Testleistungen beschrieben wurde, schnitt die gesamte Lerngruppe in einer Aufgabe unerwartet schlecht ab. Die Aufgabe passte direkt zu der derzeit im Mathematikunterricht durchgeführten Unterrichtseinheit, so dass sich das Schulleitungsmitglied 12 die Defizite nicht erklären konnte. Im unmittelbaren Anschluss an den Test erläuterte der Befragte seinen Schülern die Aufgabe und führte den korrekten Lösungsweg vor. Um zu ermitteln, inwiefern die Lernenden die Aufgabe nun verstanden haben und selbstständig bewältigen können, nahm er folgende Evaluation vor:

> *SL 12: Was ich gemacht habe dann daraus: Ich habe […] in der darauffolgenden Klassenarbeit genau die gleiche Aufgabe gestellt, nur spiegelverkehrt und zwar eben mit einer fallenden Gerade, Spiegelung an der zweiten Winkelhalbierenden. Und das Ergebnis war wiederum schlecht. Ich bin ratlos, warum das nicht gelungen ist! (LK 12 + SL 12, 00:20:37-6)*

Die Schüler bearbeiteten eine äquivalente Aufgabe mit einer abgewandelten Ausgangssituation in der Klassenarbeit. Die Items aus dem Test wurden nicht identisch übernommen,

sondern dienten lediglich als Orientierung. Die Ergebnisse zu diesen Aufgaben waren jedoch erneut defizitär. Das Schulleitungsmitglied gelangte zu der Einsicht, dass die Besprechung der Aufgabe aus der Lernstandserhebung nicht die Behebung des Defizits oder eine Weiterentwicklung der Fähigkeiten der Schüler zur Folge gehabt hatte und der erwünschte Effekt somit nicht eingetreten war. Die Gründe hierfür konnte er sich erneut nicht erklären. Weitere Maßnahmen zur Förderung ergriff er in diesem Kontext nicht.

Bei keinem weiteren Befragten, der in der Phase der Aktion konkrete Maßnahmen umgesetzt hatte, konnte eine Evaluation konnotiert werden. Demnach wurden die Handlungen vermutlich relativ unreflektiert durchgeführt und das Erreichen der Zielsetzung zur Förderung einer bestimmten Kompetenz nicht überprüft.

10 Einflussnehmende Bedingungen

In Abschnitt 5.5 wurde das Zyklenmodell für Vergleichsarbeiten nach Helmke (vgl. Helmke A. , 2004, S. 100) ausführlich dargelegt. Einer großen Bedeutung wird in diesem Modell den individuellen, schulischen und externen Bedingungen zugeschrieben, die in sehr unterschiedlichem Maße auf die Arbeit mit den Testergebnissen einwirken können.
Nach der Analyse der Nutzungsphasen folgt nun eine Betrachtung dieser Bedingungen und deren Einfluss auf die Verwendung der Lernstandserhebungen bei den untersuchten Probanden.

10.1 Individuelle Bedingungen

Bezüglich der individuellen Bedingungen wurden in der Untersuchung insbesondere motivationale Aspekte sowie der Einfluss des professionellen Selbst analysiert.

10.1.1 Intrinsische und extrinsische Motivation

Die motivationalen Aspekte äußerten sich vorrangig in der Bereitschaft, die Lernstandserhebung in den eigenen Klassen durchzuführen. Dies ist in Hessen von besonderer Relevanz, da die Teilnahme freiwillig ist und auf der individuellen Motivation beruht. Zu unterscheiden ist zwischen einer intrinsischen und einer extrinsischen Motivation. Während bei der intrinsischen Motivation ein persönlicher Wille zu der Teilnahme vorhanden ist, entsteht die extrinsische Motivation aufgrund äußerer Einflussfaktoren.
Bei einer Differenzierung zwischen den befragten Schulleitungen und den Lehrkräften ist auffällig, dass alle zwölf Schulleitungen eine intrinsische Motivation aufwiesen, währenddessen dies bei lediglich sechs Lehrpersonen festgestellt werden konnte. Dies erklärt sich aus den Teilnahmebedingungen für die Lernstandserhebungen. Jede Schule kann für selbst über eine Beteiligung entscheiden. Aus diesem Grund ist es selbstverständlich, dass die Schulleitungen bei einer Teilnahme ein Eigeninteresse aufweisen. Die Auswahl, in welchen Klassenstufen und in welchen Fächern in der einzelnen Schule die Tests durchgeführt werden, erfolgte in den betrachteten Schulen jedoch auf verschiedene Weisen. In sechs Schulen wurde den Lehrkräften freigestellt, ob sie selbst die Lernstandserhebungen durchführen möchten. Bei diesen Probanden lag folglich eine intrinsische Motivation vor, da sie sich ohne größere Beeinflussung durch das Kollegium oder der Schulleitung zur Teilnahme ent-

schlossen haben. In den anderen sechs Schulen wurde die Beteiligung der betreffenden Lehrpersonen durch die Schulleitung stark forciert beziehungsweise festgelegt, so dass in diesen Fällen eine extrinsische Motivation zu beobachten war. Im Folgenden werden sowohl die Beweggründe der intrinsischen Motivation als auch die Wahrnehmung der extrinsischen Motivation durch die Probanden dargelegt.

In Bezug auf die intrinsische Motivation konnten bei den Befragten hauptsächlich vier Begründungen festgestellt werden. Zu einen ist eine allgemeine Neugierde zu diesem Testinstrument und den Resultaten im hessischen Vergleich zu nennen.

> *LK 5: Mich hat es einfach interessiert, wie Aufgaben formuliert werden, wie die strukturiert sind, wie das organisiert wird. Das hat mich einfach wirklich interessiert. (LK 5, 00:00:20-5)*

> *LK 4: Aber grundsätzlich finde ich es eine interessante Idee, die eigenen Fortschritte, die Fortschritte auch der eigenen Klasse in einem größeren Kontext auch zu sehen und besser einzuschätzen, wie der Lernstand jetzt im Vergleich zu anderen Gruppen oder zu gleichaltrigen Gruppen aus dem selben Jahrgang steht. (LK 4, 00:00:40-8)*

Im Vordergrund standen ein allgemeines Interesse sowie das Bedürfnis nach einer Positionierung der Schülerleistungen anhand eines Ergebnisvergleichs. Dabei fand bei drei von insgesamt fünf Lehrkräften, die lediglich Neugierde als Grund anführten, keine Aktion im Nutzungsprozess statt und bei den anderen beiden Probanden konnte eine geringe Aktion diagnostiziert werden. Folglich scheinen allgemeines Interesse und Neugierde nur schwache motivationale Faktoren zu sein, die in nur geringem Maße auf den Nutzungsprozess begünstigend einwirken.

Als zweiter Beweggrund zur Teilnahme an den Test konnte die Erkenntnis, dass die Lernstandserhebungen und Bildungsstandards in der näheren Zukunft eine größere Bedeutung erhalten werden, beobachtet werden.

> *SL 9: [Wir nahmen teil] unter der Maßgabe, dass die Freiwilligkeit in absehbarer Zeit in Verpflichtung umgewandelt werden würde und wir dann gerne vorbereitet sein wollten. (SL 9, 00:00:47-3)*

Demnach erkannten diese Befragten eine Sinnhaftigkeit, sich möglichst frühzeitig mit dem Testinstrument vertraut zu machen. Dies zeugt von einer gewissen Innovationsbereitschaft und Offenheit gegenüber neuen Maßnahmen. Gleichzeitig offenbart es die Befürchtung, dass mit einer eventuell zukünftigen Verpflichtung ein zunehmender Druck einhergehe, bei den Tests positiv abschneiden und sich dahingehend positionieren zu müssen. Eine Beeinflussung dieser motivationalen Haltung auf die Nutzungsintensität der Lernstandserhebungen konnte jedoch nicht festgestellt werden.

Des Weiteren verbanden zwei der befragten Lehrkräfte mit der Teilnahme die Hoffnung, ihre eigene Professionsentwicklung in Hinblick auf die Kompetenzorientierung stärker zu fördern.

LK 1: […] [W]eil ich auch gehofft habe, ich kriege da vielleicht nochmal mehr Hilfestellung in Bezug auf kompetenzorientierte Arbeiten. (LK 1, 00:31:53-0)

LK 12: Bei der Lernstandserhebung, an der ich selbst ja das erste Mal teilnahm, war für mich […] interessant […] die Frage: Hätte ich diese Aufgabe auch den jeweiligen Kompetenzen und Anforderungsbereichen, die angegeben waren, zugeordnet? (LK 12, Nachgespräch per E-Mail)

In diesen beiden Fällen stand die Erwartung im Vordergrund, die eigene Kompetenz mithilfe der Lernstandserhebungen weiterzuentwickeln und dahingehend Anregungen zu erhalten. Zugleich ist hiermit eine unbewusste Absicherung und Bewertung der eigenen Unterrichtspraxis impliziert. Insbesondere bei der Lehrkraft 12 ist eine positive Wirkung dieser Motivation auf den Nutzungsprozess ersichtlich, indem sie eine äußerst intensive Rezeption vornahm und die Konzeption der Lernstandserhebungen analysierte. Direkte Konsequenzen auf das eigene Agieren in der Nutzungsphase der Aktion zog dies jedoch nicht nach sich.

Als vierte intrinsische Motivation wurde bei einer Schule die Passung der Zielsetzung der Lernstandserhebungen mit dem Förderkonzept der Schule genannt. Die Schule verfolgt das Ziel, Nicht-Versetzungen und Klassenwiederholungen zu vermeiden und zugleich die Abiturientenquote zu erhöhen.

SL 7: Und wir sind natürlich da auch mit dem Vorwurf konfrontiert worden, man könne das nur erreichen, wenn man das Niveau senkt. Und insofern war es für uns die Antwort, dass wir, wann immer es Vergleichstests gibt, dass wir alleine um diesen Vorwurf zu entkräften an Lernstandserhebungen, an allen Vergleichstests freiwillig teilnehmen. (SL 7, 00:01:19-7)

Die Lernstandserhebungen wurden in diesem Zusammenhang als Möglichkeit der Rechenschaft wahrgenommen, um die guten Leistungen der Schule, gemessen an den Testergebnissen, intern sowie extern zu bestätigen.

Bei den extrinsisch motivierten Befragten ist zwischen unbewusster und bewusster extrinsischer Motivation zu unterscheiden. Im ersten Fall wurde insbesondere die eigene Einstellung zu den Lernstandserhebungen durch die Kommunikation mit dem Kollegium massiv beeinflusst, wie folgende Interviewaussagen bestätigen:

LK 4: Ich habe teilgenommen, weil es bisher bei uns so Usus war, dass teilgenommen wird. Ich habe das jetzt eigentlich gar nicht am Anfang hinterfragt. (LK 4, 00:00:40-8)

SL 1: Im letzten Jahr war es so, dass […] Kollegen […] völlig entnervt waren und das wird dann natürlich oben im Lehrerzimmer konnotiert und dann macht das die Runde, wenn du eine Lernstandserhebung machst, ist das Arbeit ohne Ende und

das ist alles schlecht gemacht. Das färbt natürlich auf manche Kollegen ab. (SL 1, 00:27:21-0)

Demnach wurde die Teilnahme an den Lernstandserhebungen als selbstverständlich betrachtet, wenn ein Großteil des Kollegiums bereits seit einigen Durchläufen teilgenommen hatte. Somit wurde die Freiwilligkeit in dieser Situation und die vorhandene extrinsische Motivation nicht hinterfragt. Demgegenüber hatten negativ kommunizierte Erfahrungen eine unbewusste Abwehrwehrhaltung gegenüber dem Testinstrument zur Folge, so dass sich die eigene Bereitschaft zur Teilnahme verminderte. Die Schulleitung der Schule 1 führte dies als einen Grund an, warum zunächst nur eine geringe Zahl an Lehrkräften zur Teilnahme an den Lernstandserhebungen bereit war.

Die bewusste extrinsische Motivation wurde in allen beobachten Fällen über Druckausübung durch die Schulleitung erreicht, wie folgende Aussagen aufzeigen:

LK 2: Es ist so, dass diese Freiwilligkeit hier an unserer Schule nicht ganz so wahrgenommen wird. Der Schulleiter hat das sehr stark forciert. Ich habe die Diskussion im vorigen Jahr und vor zwei Jahren mitbekommen, wo es Widerstände aus der Fachschaft gegen diese Lernstandserhebung gab. Der Schulleitung hat sich durchgesetzt. (LK 2, 00:01:19-0)

SL 1: Wobei wir in diesem Jahr von der Schulleitung das relativ strikt eigentlich gesetzt haben, die sechsten Klassen machen das durchgängig - Schluss. (SL 1, 00:29:16-0)

Zum einen wird daraus ersichtlich, dass durch das Forcieren der Schulleitung die Teilnahme erhöht werden kann. In diesen Fällen entstand bei den Lehrkräften aufgrund der hierarchischen Beziehung zur Schulleitung die Wahrnehmung, dieser Bitte folgen zu müssen. Es lag somit lediglich eine „Scheinfreiwilligkeit“ vor. Dies kann wiederum negative Auswirkungen auf die innere Arbeitseinstellung zu den Lernstandserhebungen zur Folge haben. Bei zwei von drei betreffenden Lehrpersonen konnte ein frühzeitiger Abbruch des Nutzungsprozesses beobachtet werden, so dass keine Maßnahmen aus den Testergebnissen gezogen wurden. Bei einer anderen Schule hatte auch die Werbung der Schulleitung für eine Testteilnahme keinen Erfolg, da sich lediglich eine Lehrperson hierzu bereit erklärte. Die Schulleitung begründete dies mit der hohen Arbeitsbelastung des Kollegiums und schlussfolgerte, dass zukünftig noch mehr Werbung und Überzeugung notwendig seien.

Zum anderen legten in drei Schulen die Schulleitungen die Teilnahme an den Tests konkret fest und offerierten den Lehrpersonen keine Entscheidungsmöglichkeiten. Der Grund hierfür war die Generierung von möglichst vielen teilnehmenden Klassen. Die bietet gute Voraussetzungen für eine spätere kooperative Auswertung. Jedoch fand eine solche gemeinschaftliche Analyse in den betreffenden Schulen nicht bis nur geringfügig statt, so dass die Zielsetzung der Schulleitung verfehlt wurde.

10.1.2 Professionelles Selbst

Im Kontext des professionellen Selbst wurden zunächst die Äußerungen der Befragten auf ihre individuelle Innovationsbereitschaft hin analysiert. Vier Personen kommunizierten eine grundsätzlich negative Haltung gegenüber schulpolitischen Reformmaßnahmen, wie die folgende Interviewaussage exemplarisch aufzeigt:

> *LK 2: Dagegen was von oben, von Wiesbaden kommt oder sonst was, das sieht man mit einem dicken Fragezeichen, denn das hat oft mit der Misere hier vor Ort, mit dem Überlastetsein, mit dem Gefühl des Überfordertseins wenig zu tun. Und da werden Dinge uns auf- und übergestülpt, die wir eher mit Skepsis sehen. Vor allem weil alle paar Jahre kommt etwas Neues und das wird als letztes Evangelium dann verkündet. Und das ist problematisch. (LK 2, 00:35:04-3)*

Bei allen vier Personen war die pessimistische Einstellung persönlichen und individuellen Erfahrungen geschuldet, wie der Wahrnehmung eines zunehmenden Arbeitspensums und der Art und Weise der Implementierung von Reformen. Dies beeinflusste die allgemeine Innovationsbereitschaft im negativen Sinne, was sich auch auf die Annahme der Lernstandserhebungen als neues Leistungsmessungsinstrument niederschlug. Der Vielfältigkeit der Reformmaßnahmen im Verlauf der Dienstzeit bewirkte bei den Befragten nicht den Eindruck einer nachhaltigen Entwicklung. Vielmehr war die Wahrnehmung zu beobachten, dass die Reformen ihre Zielsetzungen verfehlen würden. Daher sank das generelle Vertrauen in deren Wirkungen, was sich auf die Lernstandserhebungen übertrug. Selbst bei einer positiven Wertung der Testkonzeption überwog bei diesen Befragten der Zweifel, inwiefern die Teilnahme tatsächlich zu einer Verbesserung der Unterrichtsentwicklung führt. Die Zweckmäßigkeit der Tests wurde äußerst skeptisch betrachtet.

Dies hatte allerdings keine negativen Konsequenzen auf die Nutzungsintensität der Ergebnisse, welche bei den vier Befragten in unterschiedlichem Ausmaß verlief. Daraus lässt sich einerseits schlussfolgern, dass die Bestätigung der vorherigen Annahme eines geringfügigen Effektes wohl erst Konsequenzen auf den nächsten Testdurchlauf haben wird. Dies zeugt von einer vorhandenen Innovationsbereitschaft unter der Voraussetzung, dass die Maßnahmen tatsächlich der Lehrkraft nutzen und nicht deren Eigenverantwortung und Gestaltungsfreiraum einengen.

Dagegen ließ sich bei anderen Befragten eine mangelnde Innovationsbereitschaft aus verschiedenen Gründen feststellen.

> *SL 6: […] wir haben genug zu tun und gerade in gewissen Phasen des Schuljahres besonders viel zu tun. Und da müssen wir uns nicht noch Zusatzbelastung aufhalsen. (SL 6, 00:34:11-4)*

SL 11: Ich muss ich mich auch durch diese Fortbildungsmaßnahmen selbst dazu zwingen, damit ich mich damit befasse. Denn von allein gucke ich da auch wieder nicht rein. (LK 11 + SL 11, 00:31:42-1)

Folglich wurde bei dem Schulleitungsmitglied 6 die Lernstandserhebung als eine erhebliche Mehrarbeit wahrgenommen. Das hatte die Entstehung einer Abwehrreaktion zur Konsequenz, so dass versucht wurde, diese Zusatzbelastung zu vermeiden oder zu minimieren. Dies führte nach Beobachtung des Schulleitungsmitgliedes 6 wiederum zu einer minimierten Nutzungsintensität der Testergebnisse im Kollegium. Das Argument der fehlenden Zeit wurde von insgesamt neun Befragten angeführt. Sie waren nicht bereit, die Grenzen ihrer Arbeitsbereitschaft zu übertreten, auch wenn ein eventueller Nutzen der Tests erkannt wurde. Das kann als eine Art Selbstschutz vor einer Arbeitsüberlastung betrachtet werden, für welche die Lehrkräfte nach eigener Anschauung keine Honorierung erhalten würden. Diese Wahrnehmung führte zu einer negativen Beeinflussung der Akzeptanz der Lernstandserhebungen, bevor die Lehrpersonen die Auswertung überhaupt begannen.

Das zweite Zitat spricht hingegen die generelle Überwindungskraft an, sich auf Innovationen einzulassen. Die eigene intrinsische Motivation sei zu gering, so dass ein konkretes Angebot oder ein externer Anstoß dazu nötig sei. Dennoch wurde in beiden Fällen die Notwendigkeit einer Weiterentwicklung der Professionalität auch in Hinblick auf die Lernstandserhebung erkannt, obwohl sie dem Test nur einen geringen Stellenwert beimaßen. Fortbildungen, die Auseinandersetzung mit neuen Konzepten sowie die Relevanz von Feedback wurden von dieser Gruppe der Befragten als Teil der eigenen Professionalität verstanden. Allerdings bestand eine Differenz zwischen dem Willen der Lehrkraft und der tatsächlichen Umsetzung, so dass die eigene Innovationsbereitschaft praktisch nicht nutzbar gemacht wurde.

Der dritte Befragte gab hingegen sein Dienstalter als Begründung an. Da er kurz vor der Pensionierung stehe, habe er kein Interesse, sich mit dieser Innovation intensiver auseinanderzusetzen. Generell konnten bei der Untersuchung keine auffälligen Zusammenhänge zwischen dem Dienstalter und der Nutzungsintensität der Lernstandserhebungen festgestellt werden.

Als positiven Einflussfaktor ist die jeweilige weiterführende Qualifizierung der Probanden zu erwähnen. Bei zwei Befragten, die nach eigener Angabe an dem Fortbildungsprojekt „Steigerung der Effizienz des mathematischen und naturwissenschaftlichen Unterrichts“ (SINUS) teilgenommen hatten, war eine intensive Rezeption und Reflexion zu beobachten. Dies kann mit der bereits vorhandenen Kenntnis zu dem Einsatz kompetenzorientierter Unterrichtsmaterialien im Fach Mathematik begründet werden. Die Lehrkräfte waren mit den

Aufgabenstellungen bereits vertraut und hatten ein größeres Interesse, deren Verwendung zu Testzwecken sowie die Ergebnisse ihrer Schüler zu untersuchen.
Des Weiteren zeichneten sich drei weitere Probanden durch eine zusätzliche Berufsqualifizierung aus, indem sie als Fachdidaktiker an Universitäten oder als Fachausbilder an Studienseminaren tätig waren. Alle drei Personen führten eine äußerst intensive Analyse der Testergebnisse sowie der Testkonzeption durch. Des Weiteren konnten bei Ihnen Maßnahmen in der Aktionsphase festgestellt werden. Durch ihre Tätigkeiten waren sie bereits mit dem Konzept der Kompetenzorientierung vertraut und setzten dies in ihrem Unterricht um. Daher empfanden sie die Lernstandserhebungen in Hinblick auf ihre eigene Unterrichtskonzeption als interessant. Eine vorige Kenntnis und Erfahrungen in der Kompetenzorientierung sind demnach in besonderem Maße für die Nutzungsintensität der Lernstandserhebungen förderlich.

10.2 Schulische Bedingungen

Zu den schulischen Bedingungen, die auf die Nutzung der Lernstandserhebungen Einfluss nehmen, sind vorrangig die vorhandenen Kooperationsstrukturen in den Schulen zu betrachten. Diese können die spätere kollektive Auswertung der Testergebnisse erheblich befördern beziehungsweise hemmen.
Bei der Hälfte der untersuchten Schulen waren keine tiefgreifenden Kommunikationsstrukturen vorhanden. Folglich fand eher ein lockerer, sporadischer Austausch zwischen den Lehrpersonen in Form von Hilfestellungen, Absprachen und Materialienaustausch statt. Zudem bezog sich die Kooperation vor allem auf eine Zusammenarbeit zwischen Kollegen innerhalb einer Jahrgangsstufe. Eine engere Teamarbeit wurde lediglich bei der Konzeption und Durchführung von schulinternen Parallelarbeiten vorgenommen. Die Kooperation war bei diesen Schulen vor allem durch individualistische Ausprägungen gekennzeichnet. Als zentrale Hindernisse für eine intensivere Zusammenarbeit wurden die Größe des Kollegiums sowie fehlende Strukturen genannt. Zudem gab es in einer Schule persönliche Konflikte innerhalb der Fachschaft, was aufgrund von Vorbehalten und fehlendem Vertrauen die Zusammenarbeit massiv erschwerte. Bei einer vergleichenden Analyse der Befragten dieser Schulen hinsichtlich der Nutzungsintensität der Lernstandserhebungen ist auffällig, dass die sechs Lehrkräfte bis auf eine Ausnahme nur geringfügig bis gar nicht die Testergebnisse kooperativ ausgewertet und reflektiert haben.
Demgegenüber waren bei den sechs anderen Schulen konkrete Kooperationsstrukturen vorhanden. Beispielsweise wurden Jahrgangskoordinatoren ernannt, welche die Zusam-

menarbeit initiierten und organisierten. In diesen Fällen wurden von allen Befragten angeführt, dass sie zusätzlich zu den Vergleichsarbeiten weitere Klassenarbeiten gemeinsam entwerfen. Dies erfordert eine hohe Kooperation, da der Unterricht inhaltlich sowie zeitlich permanent aufeinander abgestimmt werden muss und die Zusammenarbeit folglich sehr intensiv verläuft. Bei den Befragten dieser sechs Schulen konnte auch eine kooperative Auswertung der Lernstandserhebungen festgestellt werden, wenn auch in unterschiedlicher Intensität. Daher scheint sich die Etablierung von schulischen Kooperationsstrukturen auf eine eventuelle Zusammenarbeit bei der Nutzung der Testergebnisse positiv auszuwirken, was wiederum zusätzliche vergleichende Perspektiven ermöglicht. Einschränkend muss hierbei angeführt werden, dass die Äußerungen der Probanden zu schulischen Kooperationsstrukturen auf individuellen Ausprägungen beruhen. Es ist insbesondere für Lehrkräfte schwierig, Aussagen für ein Kollegium zu treffen, da sie selbst einen bestimmten Kooperationsgrad umsetzen und sich dieser nicht auf eine ganze Gruppe von Lehrkräften verallgemeinern lässt.

Spezifisch zu den Lernstandserhebungen haben einige Schulen Maßnahmen ergriffen, die sich ebenfalls in den Bereich der schulischen Bedingungen einordnen lassen. Zum einen wurde in zwei Schulen die Korrektur der Tests extern an Studenten oder im Rahmen von Arbeitsbeschaffungsmaßnahmen verlagert. Auf diese Weise sollte eine Arbeitserleichterung für die Lehrkräfte erreicht werden. Allerdings wurde die inhaltliche Auseinandersetzung mit dem Testgegenstand und den Schülerleistungen enorm erschwert, so dass die Lehrpersonen differenziertere Auswertungen der Ergebnisse nicht vornahmen. Andererseits hatte die externe Korrektur den Vorteil, dass die Lehrkräfte die Lernstandserhebung nicht in Verbindung mit einem erhöhten Mehraufwand in negativer Weise konnotiert haben. Der Nutzungsprozess der Befragten in diesen beiden Schulen brach jeweils frühzeitig ab und es konnte nur eine geringfügige Nutzung ohne Konsequenzen festgestellt werden. Dies determiniert wiederum die Behauptung, dass die fehlende Korrektur für eine intensive Auseinandersetzung mit dem Test eher hinderlich ist.

Die Schulleitungen von zwei weiteren Schulen ergriffen die Maßnahme, jeder Lehrkraft pro teilnehmende Klasse einen Korrekturtag zur Verfügung zu stellen.

> *SL 2: Denn - das muss man auch sagen- die große Akzeptanz hier im Kollegium liegt daran, dass ich auch für jede teilnehmende Klasse einen Korrekturtag frei gebe, dem Lehrer. Denn das wäre sonst noch ein Oben-drauf-Gepacke. […] Und deshalb fällt das auf fruchtbaren Boden und es kommt zu hohen Teilnehmerzahlen. (SL 2, 00:04:47-9)*

Der Korrekturtag diente einerseits einer Honorierung der geleisteten Arbeit, andererseits schaffte es für die Betroffenen Zeiträume, um die Korrektur und Ergebniseingabe vorzu-

nehmen. Diese Initiative war aus den Erfahrungen und der negativen Bewertung der Mehrarbeit durch die Kollegen entstanden, so dass einer mangelnden Akzeptanz der Lernstandserhebungen aufgrund des Zeitaufwandes vorgebeugt werden sollte. Zugleich symbolisierte diese Geste der Schulleitungen, dass sie sich von dem Testinstrument einen gewissen Nutzen erhoffen und der Durchführung und Korrektur eine Bedeutung zugesprochen wird. Das zitierte Schulleitungsmitglied 2 erhob in diesem Zusammenhang zugleich die Forderung, dass ein solcher Ausgleich nicht finanziell zu Lasten der Einzelschule sein könne, sondern vom LSA übernommen werden müsste, wenn eine flächendeckende Beteiligung an den Lernstandserhebungen anstrebt wird.

> *SL 2: […] Aber Freiwilligkeit muss dann auch auf Lehrerseite, die die Arbeit damit haben, entsprechend auch entlastet und honoriert werden. So für umsonst und Schule, sieh mal zu, wie du klar kommst, können wir nicht weitergehen. Also da muss dann etwas kommen, damit auch die Akzeptanz und die Verbreitung größer werden. (SL 2, 00:08:35-1)*

Der Korrekturtag schien bei den zwei Schulen eine positive Wirkung zu erzielen. Beide zugehörigen Lehrkräfte zeichneten sich durch eine intensive Rezeption und Reflexion sowie eine vorhandenen Aktionsphase aus.

10.3 Externe Bedingungen

Im Kontext externer Bedingungen sind vor allem die organisatorischen Rahmenbedingungen der Lernstandserhebungen von besonderem Interesse.

Zum einen wurde in diesem Zusammenhang in fünf Interviews die Freiwilligkeit der Tests von den Befragten angesprochen. Diese wurde mehrheitlich pessimistisch interpretiert, da mit dieser Freiwilligkeit keine Anerkennung des Arbeitseinsatzes einhergehe.

> *SL 10: Die Sache mit der Freiwilligkeit ist natürlich ein Problem! Die Kollegen sind belastet und die Kollegen sind also auch, besonders diejenigen, die schon etwas länger an der Schule sind, nicht so ohne Weiteres bereit, freiwillig da noch etwas zu übernehmen. (SL 10, 00:02:38-4)*

Aus dieser Äußerung eines Schulleitungsmitgliedes wird ersichtlich, dass ein möglichst geringer Arbeitsaufwand stets ein zentrales Anliegen der Lehrkräfte ist und eine Honorierung dieser Tätigkeit erwartet wird. Den Lehrenden schien nicht bewusst zu sein, dass das LSA die Rückmeldungsberichte als eine Form einer solchen Honorierung interpretiert, indem jede Lehrkraft Informationen zum Kompetenzstand ihrer Lerngruppe erhält und hieraus Konsequenzen für die Unterrichtsentwicklung ableiten kann. Die Probanden, welche die Freiwilligkeit negativ beurteilten, forderten eine Verpflichtung zur Teilnahme. Dies würde

zu einer höheren Teilnehmerzahl und zu besseren Vergleichswerten führen, ohne jedoch die Ergebnisse zwangsläufig veröffentlichen zu müssen.

> *SL 2: Aber es sollte schon im Endstadium als verbindliches Instrument an allen Schulen vorhanden sein, weil es hilft. […] [Wenn] es verbindlich gemacht wird, muss es auch etwas dafür geben. Das können die Schulleiter nicht aus ihren leeren Taschen bezahlen, denn die haben dafür keine Fonds, das ist das Problem.(SL 2, 00:50:44-3)*

Die Befragten erkannten durchaus eine Sinnhaftigkeit in den Lernstandserhebungen und konnten infolgedessen ein Verharren auf dem Prinzip der Freiwilligkeit nicht nachvollziehen. Eine Verpflichtung würde dem Aufwand der Testkonzeption und der wissenschaftlichen Auswertung gerecht werden und zugleich die Bedeutung der Lernstandserhebungen untermauern. Hiermit war zugleich die Hoffnung verknüpft, dass die Lehrkräfte ernsthafter an die Nutzung der Testergebnisse herantreten würden.

Demgegenüber befürwortete ein Schulleitungsmitglied die Freiwilligkeit der Tests, da sich die Lehrkräfte auf diese Weise bewusst für eine Teilnahme entscheiden könnten. Es entstünde eine größere intrinsische Motivation zur Arbeit mit den Tests, welche sich wiederum in einer intensiveren Nutzung äußern könne.

Zum anderen wurde der Testzeitpunkt als massiver externer Einflussfaktor auf die Nutzung von der Mehrheit der Befragten angeführt. Sie empfanden die Durchführung der Lernstandserhebung Ende Februar beziehungsweise Anfang März aus verschiedenen Gründen für nicht zweckmäßig, wie folgende Aussagen aufzeigen:

> *SL 11: Für den Unterricht würde ich mal-, also da kann ich nur für mich reden, hat es wenig Auswirkung, weil die Lernstandserhebung zum falschen Zeitpunkt kommt, das Schuljahr so gut wie abgeschlossen ist, bis ich das Detailergebnis habe, man oft die Schulklassen abgibt. Mit der Übergabe geht im Prinzip fast alles verloren oder nur das Gröbste bleibt hängen. (LK 11 + SL 11, 00:55:27-3)*

> *LK 6: […] Aber es war jetzt schon ein relativ langer Zeitraum, bis die Ergebnisse vorlagen. Also vielleicht könnte man das da noch irgendwie ein bisschen beschleunigen. Das wäre ganz schön, dass auch die Rückmeldung für die Schüler noch zeitnaher erfolgt. Weil, ich denke, den meisten ist das gar nicht mehr so präsent. (LK 6, 00:29:39-1)*

Aufgrund der Platzierung der Lernstandserhebungen im zweiten Schulhalbjahr nahmen die Lehrkräfte vor allem den Lehrplandruck und die zusätzliche Arbeitsbelastung durch das Abitur wahr, so dass sie aufgrund der zeitlichen Enge eine intensivere Auseinandersetzung mit der Lernstandserhebung vermieden. Verschärft wurde dies durch die Zeitspanne zwischen der Testdurchführung und dem Erhalt des ausführlichen Rückmeldeberichts, welche als zu lang empfunden wurde. Die Lernstandserhebungen seien sowohl den Schülern als auch den Lehrpersonen nicht mehr präsent gewesen. Eine langfristige Nutzung der Testergebnisse konnte demnach nicht mehr stattfinden, denn es verblieben nur wenige Unter-

richtswochen. Im folgenden Schuljahr wurden die Klassen oftmals neu zusammengesetzt oder von anderen Lehrkräften unterrichtet. Die Informationen müssten somit der neuen Lehrperson weitervermittelt werden, die sich jedoch im ungünstigen Fall nicht mit den Lernstandserhebungen auseinandergesetzt hat und somit auch die Ergebnisse nicht interpretieren kann.

Die Probanden forderten daher einen Testzeitpunkt, welcher genügend Zeit für eine Nutzung der Ergebnisse im laufenden Schuljahr ermöglicht.

> *LK 3: Aber je mehr ich darüber nachdenke, desto besser gefällt mir die Idee, dass man diese Lernstandserhebung am Anfang des Schuljahres macht. Man will ja auch nicht nachprüfen, was ist im Kurzzeitgedächtnis hängengeblieben, sondern was ist wirklich da? (LK 3, 00:28:07-8)*

Der Vorschlag, die Lernstandserhebung zu Beginn eines Schuljahres durchzuführen, wurde mehrfach erwähnt. Dies würde noch immer der Anforderung nach einer Messung von nachhaltigen Kompetenzen gerecht werden, hätte allerdings eine Verlagerung der Schwerpunktsetzung des Tests zur Folge. Die Funktion einer Selbstreflexion durch die Lehrkraft träte erheblich zurück, weil diese die Klasse oftmals neu übernommen hätte und keine Rückschlüsse auf ihr eigenes Agieren ziehen könnte. Hingegen würde der Test ein größeres diagnostisches Gewicht erhalten, da die Lernausgangslage ermittelt würde und darauf aufbauend Schlussfolgerungen für die weitere Unterrichtsarbeit gezogen werden könnten.

Zusätzlich belastend wurde die Dopplung von externen Tests im Fach Mathematik wahrgenommen. Insbesondere in der Klassenstufe 8 findet neben den Lernstandserhebungen auch der Mathematikwettbewerb statt.

> *SL 3: Also wir haben in der Jahrgangsstufe 8 den Mathematikwettbewerb, der verpflichtend ist als Klassenarbeit und wir haben die Lernstandserhebungen, die noch freiwillig sind. Aber wir haben einfach zwei […] von außen gesetzte Tests. Das ist nicht besonders sinnvoll! […] So und was […] ich besonders bemängele ist an diesen zwei Tests dann, dass die beiden so gar nichts miteinander zu tun haben. (SL 3, 00:28:52-9)*

Die Zielsetzungen der beiden Testformate sind sehr verschieden, so dass insgesamt drei Befragte nicht den Sinn darin erkennen konnten, warum beide Tests weiterhin parallel bestehen bleiben. Es existiert keine gemeinsame Ausrichtung in Hinblick auf die Kompetenzorientierung. Daher entstand sowohl bei den Lehrkräften als auch bei den Schülern ein gewisser Überdruss an extern konzipierten Arbeiten. Als Alternative wurde eine Kombination beider Testformate vorgeschlagen. Allerdings steht die Konzeption der Lernstandserhebungen konträr zu dem Charakter des Mathematikwettbewerbs, welcher benotet wird und dessen Ergebnisse veröffentlicht werden.

11 Bewertung der Lernstandserhebungen durch die Lehrkräfte und Schulleitungsmitglieder

Nachdem sowohl die Nutzung der Lernstandserhebungen durch die Befragten als auch die Einflussnahme von weiteren Bedingungen analysiert worden sind, steht nun die Bewertung der Lernstandserhebungen durch die beteiligten Personen im Zentrum der Betrachtungen. Alle Probanden sprachen diesen Aspekt selbst an und legten ihre Wahrnehmung diesbezüglich ausführlich dar. Daher scheint die Bewertung der Tests ein bedeutender Faktor zu sein, welcher auf die Nutzungsintensität einwirkt. Die Äußerungen der Befragten lassen sich sechs thematischen Bereichen zuordnen:

- Bewertung der Testdurchführung,
- Bewertung der Testinhalte und -schwierigkeit,
- Bewertung der Testkorrektur,
- Bewertung der Rückmeldeberichte,
- Diskussion über eine Benotung der Testresultate sowie
- Bewertung des persönlichen Nutzens der Lernstandserhebung.

Im Folgenden werden diese einzelnen Bewertungskategorien ausgewertet.

11.1 Bewertung der Testdurchführung

Bezüglich der Rahmenbedingungen bei der Durchführung der Lernstandserhebungen erwähnten einzelne Befragte spezifische Kritikpunkte, die sie aufgrund ihrer individuellen Erfahrung als negativ werteten. Zum einen beklagte beispielsweise ein Schulleitungsvertreter, dass der organisatorische Aufwand für den Schulkoordinator zu groß sei. Ein anderer Proband führte die Unsinnigkeit der Bücherfrage an, dessen Fragebogen die Schüler bei jeder Lernstandserhebung erneut ausfüllen müssen. In einem anderen Fall wurden anfangs falsche Materialien an die Schule gesandt. Die genannten Aspekte stellen Einzelmeinungen dar.

Häufiger wurde hingegen von insgesamt sieben Befragten der Umfang sowie die für die Bearbeitung zur Verfügung stehende Zeit beurteilt, wie folgende Interviewaussagen exemplarisch aufzeigen.

LK 4: Wenn ich an den Englischtest denke, ist mir dieser hohe Umfang gleich präsent […]. Die Texte waren ja nicht gerade klein. Die Hörverstehenstexte […] waren sehr schnell vorgetragen. Und die Schüler hatten auch wenig Zeit, über das, was sie da produzieren, großartig nachzudenken. (LK 4, 00:16:01-3)

LK 8: Da fand ich [im Reading Comprehension] sieben Texte viel zu viel. […] Dazu war die Zeit viel zu kurz. Also meiner Meinung nach hätten es weniger Texte sein können. […] Und auf der anderen Seite das Writing: Da hatten Sie meiner Meinung nach viel zu viel Zeit. Da waren vierzig Minuten, da war meine gesamte Klasse innerhalb von zwanzig Minuten fertig. (LK 8, 00:04:17-3)

SL 4: Und das andere war einfach unglaublich viel, unglaublich vielfältig und ich denke, die mussten durchhalten. […] Und das Gleiche haben mir die Deutschkollegen zurückgemeldet […]. Und die haben gesagt: "Das war eigentlich-, wollen wir testen, wie schnell die Schüler lesen, wie viel Durchhaltevermögen sie haben, wie die sich konzentrieren können oder hat es etwas mit Deutschunterricht zu tun?" (SL 4, 00:08:50-8)

Demnach wurde der Umfang der Lernstandserhebungen als problematisch und teilweise als unpädagogisch wahrgenommen. Die Schüler reagierten nach der Beobachtung der Befragten auf die Testfülle mit Unsicherheit und teilweise mit Panik. Sie fühlten sich unter Zeitdruck gesetzt. Insbesondere in der Klassenstufe 6 wurde eine zeitliche Überforderung der Lernenden wahrgenommen, da die Schüler es noch nicht gewöhnt seien, Arbeiten über 90 Minuten zu schreiben. Dies hatte Folgen für die Qualität der Leistungen. Indem die Schüler nicht alle Aufgaben bearbeiten beziehungsweise ihre Antworten nicht überprüfen konnten, entstand ein größeres Fehlerpotenzial und es wurden Items mit null Punkten bewertet. Die Resultate spiegelten somit nicht die tatsächliche Leistungsfähigkeit der Schüler wieder. Zwei Befragte empfanden diesen Umstand als so kontraproduktiv, dass sie die tatsächlich erreichte Zielsetzung der Tests infrage stellten. Es seien primär nicht fachliche Fähigkeiten der Lernenden getestet worden, sondern wie schnell sie unter Druck arbeiten können.

Zudem wurde von zwei Probanden angemerkt, dass ihrer Einschätzung nach der Umfang der Lernstandserhebung unnötig überfüllt sei, da mehrere Aufgaben die gleichen Kompetenzen erfassen würden. Es würde daher ausreichen, lediglich eines dieser Items zu verwenden und den Testumfang zu reduzieren. Des Weiteren wurden die festgesetzten Zeitvorgaben zur Bearbeitung einzelner Testbereiche kritisiert. Diese seien unausgewogen gewesen und hätten nicht dem sonstigen Vorgehen im Unterricht entsprochen.

11.2 Bewertung der Testinhalte und -schwierigkeit

Ein bedeutsamer Aspekt für die Bewertung der Lernstandserhebungen stellte für die Befragten die Lehrplanadäquatheit sowie die inhaltliche Passung zu dem eigenen Unterricht dar.

SL 8: Also ich denke, das sind schon so die Bereiche, die Kompetenzen, die halt auch vorhanden sein sollten in dem Alter. (SL 8, 00:08:11-7)

SL 9: Das sind sehr auf funktionale Texte ausgerichtete Textgrundlagen, sehr funktionale Texte. Also der ganze Bereich, den wir ja auch in der Mittelstufe versuchen aufzubauen, literarische Bildung et cetera, der spielt gar keine Rolle. (SL 9, 00:19:59-1)

LK 3: […] Also es war so, dass da Aufgaben gestellt waren über Themen, die hatten wir im Unterricht einfach noch nicht besprochen. […] Das hatten die Kinder noch nicht geübt und da kam eigentlich auch nur Käse raus. […] Aber welche psychologisch verheerenden Folgen das für die Kinder hat! (LK 3, 00:06:25-7)

Insgesamt bestätigten acht Befragte, dass die Testinhalte grundsätzlich den Themen des Lehrplans und ihres eigenen Unterrichts entsprechen würden. Positiv wurde insbesondere die Breite der angesprochenen fachlichen Fähigkeiten gewürdigt. Trotz dieser Bewertung kritisierten sechs Probanden, dass es dennoch zentrale unterrichtliche Themen und Kompetenzbereiche gäbe, die in den Lernstandserhebungen zu wenig oder gar keinen Raum einnähmen. Die Tests verloren für die Lehrkräfte damit an Wert, da bedeutsame Unterrichtsschwerpunkte keinen Testgegenstand darstellten. Zudem wurde angemerkt, dass in den Tests Inhalte vorgekommen seien, die zu dem Testzeitpunkt noch nicht im Unterricht thematisiert worden waren. Die Schüler hatten in diesem Fall nur eine geringe Chance, die Items erfolgreich zu bewältigen. Zudem hätte dies zu einer enormen Verunsicherung bei den Schülern während des Tests geführt.

Bezüglich der Gestaltung und Konzeption der einzelnen Aufgaben wurden ebenfalls diverse kritische Aspekte genannt:

LK 1: Es geht ja in der sechsten Klasse gerade darum, ob die dieses Level A1 des Europäischen Referenzrahmens oder A2 haben. […] Es ging zum Beispiel darum eine Postkarte zu schreiben […]. Und der Unterschied zu A1, A2 ist vor allem eben auch, dass ich komplexere Sätze bilden kann, dass ich sentence connectors, linking words benutze und das tut man hier auf einer Postkarte nicht. […]. Also insofern fanden wir diese beiden Aufgabentypen also eher ungeeignet um das festzustellen. (LK 1, 00:04:58-1)

LK 4: […] Man sollte einerseits hören und verstehen, aber andererseits auch schon kurze Phrasen produzieren. Und meiner Meinung nach erzeugt das bei den Schülern da Indifferenzen, da sie dann zu Sprachproduktion auch wieder geleitet werden, was ja in dem Fall eher möglicherweise hinderlich ist. (LK 4, 00:07:14-2)

Einerseits wurde von mehreren Befragten eine Differenz zwischen dem Item und seiner deklarierten Zielsetzung festgestellt, indem die Aufgabe zum Testen bestimmter Kompetenzen ungeeignet sei oder andere Fähigkeiten vorrangig überprüfe. Andererseits hätten einige Items mehrere Kompetenzen gleichzeitig getestet und seien somit nicht eindeutig gewesen. Dies hätte wiederum zu einer kognitiven Überforderung der Schüler geführt. Weiterhin gab ein Proband an, dass einige Aufgaben unverständlich und doppeldeutig formuliert gewesen seien. Eine andere Aussage thematisierte die mangelnde Passgenauigkeit der Aufgabenformate zu denen der zentralen Abschlussprüfungen. Für eine langfristige Vorbereitung auf das Abitur sei es sinnvoll, die Aufgabenformate einander anzupassen.

Die in den Items dargestellten Kontexte bewerteten die Mehrheit der Befragten positiv. Demnach seien sie interessant und ansprechend gestaltet. Ein Proband kritisierte in diesem Zusammenhang, dass sich die Schüler oftmals in die Gedankenwelt der Erwachsenen hineinversetzen müssten. Dies sei eine nicht-kindgerechte Perspektive. Zudem seien oftmals alltagsferne Situationen dargestellt, die der realen Lebenswelt der Schüler nicht entsprächen.

Ähnlich differenziert fiel das Urteil der Befragten zu der Testschwierigkeit aus:

> *SL 7: Vom Schwierigkeitsgrad her durchaus machbar, glaube ich, für unsere Schüler. Natürlich gibt es da immer einzelne Dinge, die irgendwie schwierig sind, die von den Schülern auch als schwer empfunden werden. Andere sind eher einfach. (SL 7, 00:06:49-2)*
>
> *LK 11: Beim Schreiben selbst- (lacht), da haben sie primitive Dinge hingeschrieben. Sie waren nicht gefordert. Da mussten nicht vorkommen Connectives, da musste nicht vorkommen Gerundium, da muss kein If-Satz vorkommen, da muss ja nichts vorkommen! (LK 11 + SL 11, 00:21:42-5)*

Neun Befragte äußerten sich positiv zu der Testschwierigkeit. Demnach seien die Aufgaben grundsätzlich adäquat und angemessen gewesen. Demgegenüber empfanden drei Befragte die Lernstandserhebungen als zu leicht und forderten einen höheren Anspruchsgrad, während wiederum zwei Befragte den Test als zu anspruchsvoll beurteilten, da zum Beispiel für die Schüler neue Textsorten und unbekannte Vokabeln vorgekommen seien. Angemerkt wurde des Weiteren die Unausgewogenheit der Aufgabenschwierigkeiten. Einige Bereiche seien sehr leicht, andere wiederum zu schwierig. Diese Aussage kann auch ein Indiz sein, dass den Probanden die testtheoretischen Grundlagen bezüglich der Aufgabenschwierigkeiten unbekannt waren (vgl. Abschnitt 4.3.3).

Die negative Wertung der Tests hatte in einem hohen Ausmaß Auswirkungen auf die Akzeptanz der Lernstandserhebungen bei den schulischen Akteuren. Entsprach die Leistungsmessung nicht dem erwarteten Schwierigkeitsgrad, wurde der Test als nicht geeignet für den eigenen Unterricht eingeschätzt. Bei einer Schule führte diese Bewertung zu einer Verwei-

gerung der künftigen Teilnahme an den Lernstandserhebungen, da sie zu anspruchslos und dem Niveau der Schule nicht entsprechend seien.

Die Bewertung der Aufgabenschwierigkeit hing davon ab, was sich die Lehrkraft von den Lernstandserhebungen versprach. Hatte sie an einer Kompetenzdiagnostik für die weitere Unterrichtsentwicklung Interesse, beurteilte sie den Test tendenziell als zu leicht und forderte ein höheres Anforderungsniveau. Stand hingegen die Positionierung ihrer Lerngruppe in ihrem Fokus, war sie an einem möglichst guten Abschneiden der Schüler interessiert. In diesem Fall wurden einzelne Items der Lernstandserhebung oftmals als zu schwierig kritisiert.

11.3 Bewertung der Testkorrektur

Im Anschluss an die Durchführung der Lernstandserhebungen mussten die Fachlehrkräfte die Schülerantworten entsprechend den Lösungshinweisen korrigieren und die Ergebnisse in das Online-Portal eingeben. Hierfür erhielten sie eine detaillierte Korrekturanweisung, welche sie zu befolgen hatten. Allerdings schien die konsequente Umsetzung dieser Anweisung je nach Aufgabenformat und Unterrichtsfach den Lehrpersonen Schwierigkeiten bereitet zu haben, wie folgende Interviewaussagen belegen:

> *LK 4: Da muss ich sagen, also im Mathematikbereich war die Bewertung ja sehr eindeutig und die Korrektur auch einfach umzusetzen [..]. (LK 4, 00:22:28-7)*
>
> *LK 4: Die Beurteilung nach den Referenzstufen bzw. nach den GER-Richtlinien oder Kompetenzleveln A1 und A2 ist mir schwergefallen, weil die Referenzbeispiele gefehlt haben. Das ist ja eine sehr abstrakte […] oder allgemein gehaltene Erklärungen dieses Sprachniveaus vorhanden. Und man braucht doch irgendwo nochmal Vergleiche, weil ich kann mir vorstellen, dass das von Lehrkraft zu Lehrkraft unterschiedlich wahrgenommen wird. Wann hört A1 auf, wann beginnt A2? (LK 4, 00:12:04-7)*

Demnach war bei geschlossenen und halboffenen Items die Korrektur leicht, da lediglich ein Vergleich zwischen der kurzen Antwort und der Lösungsangabe in der Korrekturanweisung nötig war. Weil insbesondere im Fach Mathematik eine erhöhte Anzahl solcher Items vorlag, wurde das Korrekturprozedere von den Mathematiklehrkräften tendenziell positiver bewertet.

Größere Schwierigkeiten entstanden hingegen bei Items, die ausführlichere Antworten im Sinne einer Textproduktion erforderten. In der ersten Fremdsprache mussten beispielsweise die Schülerantworten den GER-Niveaustufen zugewiesen werden. Die Korrekturbeispiele fehlten oder waren nach Ansicht der Befragten nicht hilfreich, weil sie sich nicht auf die Schülerantworten übertragen ließen. Aufgrund der Schwierigkeit, die Anleitung korrekt zu

interpretieren und gemäß den Anforderungen die Punkte zu vergeben, entstand bei den Lehrpersonen eine größere Unsicherheit. Dies kostete insbesondere Zeit, da die Lehrpersonen die Antworten tiefgründiger prüfen mussten und die Korrektur folglich eine längere Zeit in Anspruch nahm. Ein Befragter führte an, dass es eigentlich notwendig sei, im Kollegium die Korrekturanweisung gemeinsam zu besprechen, um zu eventuellen Schwierigkeiten eine gemeinsame Lösung zu finden.

Diese Schwierigkeiten und Unsicherheiten hatten zur Folge, dass die Lehrerenden die Aussagekraft der ausgewerteten Testergebnisse anzweifelten.

> *LK 11: Es geht nicht um Konkurrenz, […] sondern korrigieren wir wirklich genau gleich, obwohl die Anweisungen da sind? Oder ist mein Spielraum so groß, dass ich sagen kann, das finde ich gut, das finde ich nicht gut. Wie ich andere Arbeiten auch korrigiere. Also es hängt ein bisschen von mir ab, was dabei rauskommt. (LK 11 + SL 11, 00:05:22-4)*

Indem aufgrund der Unsicherheit bei der Anwendung der Korrekturhinweise individuelle Entscheidungen zur Punktvergabe getroffen wurden, vergrößerte sich die Subjektivität in der Korrektur. Diese verminderte Objektivität der Ergebnisse erschwerte wiederum eine Vergleichbarkeit mit Parallelklassen oder mit dem hessischen Landesmittelwert. Die Lehrkräfte nahmen äußerst sensibel wahr, dass das Ergebnis ihrer Lerngruppe in einem enormen Maße davon abhängt, wie strikt sie sich an die Korrekturanweisung gehalten haben. Aufgrund dessen wurde den Aussagen der Rückmeldeberichte ein geringeres Vertrauen in Hinblick auf ihre tatsächliche Aussagekraft entgegengebracht.

Weiterhin konnte bei einem Befragten festgestellt werden, dass er sich großzügig Freiräume in der Korrektur schuf und erheblich von den Vorgaben abwich.

> *LK 1: Es gab eine Wegbeschreibung, da haben wir gesagt, die lassen wir wegfallen, weil die hatten wir noch nie gemacht. (LK 1, 00:04:58-1)*
>
> *LK 1: Der hier hat dann abgebrochen, der war schon vorher schlecht und deshalb habe ich das gar nicht ausgewertet, weil das zweite Heft hat er gar nicht mehr bearbeitet. (LK 1, 00:15:52-2)*

Die Lehrkraft hat einzelne Aufgabenteile nicht bewertet. Da diese in der Ergebniseingabe mit null Punkten gewertet werden, wirkte sich diese Korrekturentscheidung zum einen massiv negativ auf das Gesamtergebnis der Klasse aus. Zum anderen hat die Lehrperson einen Schüler aufgrund einer schlechten Leistung nicht in der Gesamtbewertung berücksichtigt. Dies führte erneut zu einer Verfälschung und Manipulation der Ergebnisse.

Neben Problemen bei der Ausführung der Korrekturanweisung übten einige Probanden Kritik an dem pädagogischen Gehalt dieser Vorgaben.

LK 2: Und da musste ich ihnen dann sagen, die Korrekturanweisungen sind sehr eng, sie sind zum Teil unpädagogisch, wie ich finde, sie sind zum Teil auch nicht sachgerecht […]. (LK 2, 00:04:34-5)

LK 4: Weil ein Schüler, der eine Aufgabe nicht bearbeitet hat, wird genauso behandelt, wie ein Schüler, der sie bearbeitet hat, aber alles falsch hatte, null Punkte erreicht. […] Was mich aber so ein bisschen gestört hat, ist dieser Fokus auf das Produkt. Das Endergebnis muss stimmen. Das ist ein Alles oder Nichts. Der Prozess bleibt eigentlich völlig unberücksichtigt. (LK 4, 00:16:01-3)

LK 2: Dieses enge Verständnis eines Textes ist falsch, zumindest problematisch! […] Die Differenz zwischen Erzählzeit […] einerseits und der „erzählten Zeit" […] wird bei dieser Lösungsvorgabe nicht beachtet. Dieser grundsätzliche Fehler in der Bewertung ließe sich noch an zahlreichen weiteren Beispielen zeigen! (LK 2, Bericht)

LK 6: Andererseits muss man dann auch sagen, sind die Kritikpunkte einmal darin zu sehen, dass Rechtschreibung und Zeichensetzung oder so die sprachlichen Aspekte des Faches Deutsch kaum bis gar nicht berücksichtigt werden. (LK 6, 00:32:40-2)

Ein Vorwurf bestand in der angeblich unpädagogischen Ausrichtung der Korrekturanweisungen. Hierzu zählt unter anderem die rigide Festlegung der Schülerantworten in Richtig oder Falsch. Teilrichtige oder unvollständige Antworten wurden als falsch gewertet, was der schulischen Praxis deutlich widerspräche. Eine differenziertere Korrektur wäre nach Ansicht der Befragten sinnvoller, zumal den Schülern eine solche Bewertung nur schwer kommunizierbar sei. Wenn der Lernende einen richtigen Lösungsansatz hat, wird seine Antwort genauso mit null Punkten gezählt wie bei jemanden, der die Aufgabe nicht bearbeitet hat. Mit dieser Vorgehensweise waren insgesamt fünf Befragte nicht einverstanden, da sie der individuellen Leistung eines Schülers nicht gerecht werde.

Des Weiteren entdeckten zwei Lehrkräfte fachliche Fehler in der Korrekturanweisung. In diesen Fällen entstand für die Lehrkräfte wiederum eine Entscheidungssituation, ob sie streng nach Vorgabe korrigieren oder nach ihren eigenen fachlichen Grundsätzen die Antworten bewerten. Dies führte zu einer verminderten Objektivität bei der Korrektur. Des Weiteren wurde von den Probanden bemängelt, dass zentrale Kompetenzbereiche, wie das sichere Beherrschen von Orthografie, Grammatik sowie die Ausdrucksfähigkeit insbesondere in den sprachlichen Fächern nur unzureichend berücksichtigt würden. Dies stehe wiederum im Gegensatz zu der herkömmlichen Bewertungspraxis. Den Schülern sei es ebenfalls unklar gewesen, warum diese Aspekte keine Gewichtung fanden. Die Lehrkräfte hatten zudem Skrupel, einen nicht nach Rechtschreibung korrigierten Text dem Schüler zurückzugeben, obwohl dieser gravierende Fehler beinhaltete.

Neben der Thematisierung der Korrekturanweisung führten sechs Befragte erneut den enormen Zeitaufwand an, der mit der Bewertung der Schülerantworten verbunden sei.

SL 12: Ich denke, der Korrekturaufwand ist erheblich. Ich habe mir wirklich aus dem Grunde, weil die Kollegen kritisiert haben, es wäre also ein riesiger Arbeitsaufwand, habe ich mir das auch mal notiert. Und ich habe definitiv mehr dafür aufwenden müssen als für die Korrektur einer normalen Klassenarbeit. (LK 12 + SL 12, 00:05:57-7)

LK 8: Und ganz schlimm fand ich halt die Auswertung, also diese Listen schreiben und dann diese ganzen Listen nochmal im Computer eingeben […] Also da müsste man, meine ich, ein anderes Schema finden, dass man das irgendwie auf Bögen macht und anstreichen und die werden dann irgendwo zentral ausgewertet. Also da finde ich mich als Lehrer ein bisschen überqualifiziert für solche Sachen. (LK 8, 00:08:43-9)

Einige Probanden empfanden die Korrekturzeit als unverhältnismäßig hoch, so dass es für sie fraglich war, ob der Aufwand dem Nutzen letztendlich gerecht werde (vgl. Abschnitt 11.6). Insbesondere die Ergebniseingabe im Online-Portal hat eine negative Bewertung hervorgerufen, da diese Arbeit keinen fachlichen Anspruch hatte und als langwierig wahrgenommen wurde. Es wurde mehrfach der Vorschlag angebracht, zumindest diesen Teil der Korrektur den Lehrkräften abzunehmen und die Ergebnisse computergestützt einzulesen. Andere Befragten beurteilten den Aufwand der Korrektur nicht als problematisch, so dass diese Bewertung anscheinend sowohl von dem Unterrichtsfach als auch von der persönlichen Arbeitsbereitschaft abhängt.

Zuletzt wurden von einigen Befragten technische Probleme bei der Korrektur und Ergebniseingabe angeführt. Beispielsweise gab es Schwierigkeiten, wenn ein Kind nach der Anmeldung zur Teilnahme an den Tests die Klasse verlassen hatte. In diesem Fall musste dennoch eine Wertung oder der Status „fehlend“ eingetragen werden, um die Ergebniseingabe erfolgreich abschließen zu können. In einem anderen Fall war es nicht möglich, eine größere Schülerzahl als fehlend zu vermerken. Die Schüler wurden mit null Punkten gewertet, was zu einer massiven Verfälschung der Ergebnisse führte. Des Weiteren wurde von einer Lehrkraft die Korrekturzeit von drei bis vier Wochen kritisiert. Durch diesen zeitlichen Druck habe sie sich mit den Lernstandserhebungen nicht tiefgründig auseinandersetzen können.

11.4 Bewertung der Rückmeldeberichte

Die Äußerungen der Befragten bezüglich einer Bewertung der Rückmeldeberichte umfassten zwei Bereiche: Einerseits betrafen sie die Konkretisierung der erhaltenen Informationen und andererseits die fehlende individuelle Auswertung für einzelne Schüler.

LK 2: Die Rückmeldung wäre noch deutlich brauchbarer, wenn die den einzelnen Aufgabenformaten zugrundeliegenden Kompetenzen und Standards explizit genannt worden wären, so das für die zukünftige Unterrichtarbeit auch ange-

messene Schlussfolgerungen gezogen werden könnten. […] Unverständlich bleibt, weshalb man auf die Bildungsstandards Deutsch der KMK von 2004 (sie sind doch auf das Ende von Sek I bezogen!) verweist, ohne dass die aktuellen Bildungsstandards Deutsch für Hessen auch nur erwähnt werden. (LK 2, Bericht)

LK 8: Und da fehlt mir so ein bisschen die Handhabung, woran kann ich da arbeiten, um da wirklich die eigentliche Diagnose der einzelnen Schüler zu verbessern. Also da müssten die Lernstandserhebungen meiner Meinung nach ein bisschen individueller sein. (LK 8, 00:14:30-5)

SL 3: Also was zum Beispiel toll wäre, ist […] so dieses Feedback für den einzelnen Schüler. […] Das fände ich dann noch sinnvoller. Weil diese Rückmeldung "Wir als Klasse"-, also ein Kollektiv ist man dann ja doch nicht (lacht), man will ja für sich selbst das wissen. (SL 3, 00:27:20-8)

Das zuerst angeführte Zitat bezieht sich auf die Konkretisierung der Rückmeldeberichte. Die Lehrkraft vermisste eine kurze Erläuterung zu den Zielsetzungen der Aufgaben in Bezug auf die getesteten Kompetenzen und angesprochenen Bildungsstandards. Sie hatte sich somit während der Korrektur und Analyse der Testaufgabe weiterführende Informationen gewünscht. Anscheinend wusste sie nicht, dass es zusätzliche didaktische Materialien gibt, in welchen diese Informationen dargelegt werden. Zugleich offenbarte der Gesprächsausschnitt, dass der Befragte einen Zusammenhang zwischen der Testkonzeption, den Rückmeldeberichten und dem hessischen Kerncurriculum vermisste. In der Tat ist es unverständlich, warum sich die Rückmeldeinformationen auf die KMK-Bildungsstandards beziehen, währenddessen die konkreteren hessischen Kompetenzerwartungen spezifisch auf die Klassenstufen 6 und 8 ausgerichtet sind.

Die anderen beiden exemplarisch angeführten Äußerungen betreffen die mangelnde Individualisierung der Rückmeldung für einzelne Schüler. Da die Informationen nicht konkret auf einzelne Schüler bezogen waren, fiel es den Lehrkräften schwer, Hinweise für die Weiterarbeit und zur spezifischen Förderung abzuleiten. Hierfür sollten die Defizite stärker herausgestellt werden. Auch für die Schüler selbst sei eine individualisierte Rückmeldung sinnvoll, um ihren eigenen Leistungsstand beurteilen zu können. Insbesondere bei neuen Klassenzusammensetzungen sei es hilfreich, für einzelne Schüler Daten und Informationen zu erhalten, die in der Schülerakte hinterlegt oder der weiterführenden Lehrperson übergeben werden könnten.

11.5 Diskussion über eine Benotung der Testresultate

Bei der Analyse der Nutzung der Lernstandserhebung durch die Befragten wurde bereits offensichtlich, dass zu der Thematik „Benotung der Testergebnisse“ uneinheitliche Positionen vertreten wurden (vgl. Abschnitt 9.3.1). Diese werden im Folgenden detailliert dargelegt.

> *LK 3: Aber diese Lernstandserhebung, die sollte ganz abgetrennt von einer Benotung sein. Die sollten uns und der Klasse zeigen, woran wir noch arbeiten müssen. Aber das sollte mal ein Bereich sein, wo man nicht schon wieder in Notenkategorien denkt. (LK 3, 00:11:29-1)*
>
> *LK 2: Also wir Lehrer sind auch daran interessiert, dass wir Arbeit nicht doppelt oder umsonst machen müssen. […] Wenn wir aber die Möglichkeit hätten, das auch zu benutzen im Sinne von Leistungsüberprüfung, wenn wir also einen Mehrwert stärker entdecken könnten, wäre das bestimmt günstiger. (LK 2, 00:29:04-0)*
>
> *SL 6: Also ich verspüre so den Wunsch bei den Kollegen, dass sie eigentlich sagen: "Warum wird das nicht mal eine Klassenarbeit?" […] Davon versprechen sich einige etwas, vielleicht auch noch eine stärkere Beteiligung und Mitarbeit der Schüler […]. (SL 6, 00:23:48-3)*

Einerseits lehnten einige Probanden eine Benotung der Testleistung radikal ab. Die Lernstandserhebungen sollten lediglich Hinweise für die weitere Arbeit liefern und in diesem Sinne als formative Kontrolle dienen. Eine Benotung wäre für weitere Fördermaßnahmen eher kontraproduktiv. Andererseits kommunizierten neun Befragte das Bedürfnis, auf diese Weise die geleistete Arbeit der Schüler als auch den eigenen Korrekturaufwand honorieren zu wollen. Die Lernstandserhebung erhielte einen Mehrwert, da sie keine zusätzliche Korrektur implizieren würde. Des Weiteren würde der Nutzung der Lernstandserhebung eine größere Bedeutung zuteilwerden. Die Befragten erhofften sich zugleich, mithilfe des Notendrucks eine stärkere Ernsthaftigkeit in der Testbearbeitung bei den Schülern und dementsprechend eine höhere Gewichtung der Lernstandserhebung zu erreichen. Als Kompromiss wurde von zwei Schulleitungsvertretern vorgeschlagen, lediglich positive Leistungen in Form einer Benotung zu würdigen, so dass die Schüler dennoch zur Testbearbeitung motiviert sind.

11.6 Bewertung des persönlichen Nutzens der Lernstandserhebung

Abschließend soll in Bezug auf die Bewertung der Lernstandserhebungen durch die Befragten der empfundene Nutzen für die eigene Unterrichtsarbeit analysiert werden. Hierzu wurden äußerst viele Aspekte der Lernstandserhebungen angeführt, die als nützlich wahrgenommen wurden. Nachfolgend werden diese in Aufzählungsform und mit exemplarischen Interviewzitaten ergänzt.

- Informationswert der Lernstandserhebungen für Diagnostik und Positionierung

  LK 2: Also einmal, sie gibt mir Aufschluss über die Leistungsfähigkeit meiner Klasse in bestimmten Bereichen, soweit sie für mich erkennbar sind kompetenzmäßig, im Vergleich zum Landesschnitt. Also ich weiß jetzt ungefähr, wo meine Klasse steht und das ist eine wichtige Information. […] Sie gibt auch für mich Aufschluss über das überraschende Ergebnis bei einzelnen Schülern, das besser bzw. schlechter ist als mein Bild, das ich gewöhnlich von ihnen entwickelt habe. (LK 2, 00:55:56-0)

  LK 7: Dass man anhand der Ergebnisse sieht, wo die Klasse Stärken und Schwächen hat und dass man insbesondere bei den Schwächen ansetzen kann, um diese Themen dann in der Klasse zu fördern, damit diese Schwächen einfach ausgemerzt werden. (LK 7, 00:43:13-9)

  Der diagnostische Gehalt der Informationen aus den Rückmeldeberichten wurde als Nutzen erkannt. Dies bot Hinweise auf die Defizite in Form einer Stärken-Schwächen-Analyse, so dass tendenzielle Rückschlüsse auf die Weiterarbeit gezogen werden konnten. Des Weiteren ermöglichte die Lernstandserhebung eine Positionierung der Leistungsstärke der Lerngruppe im Vergleich zu dem hessischen Landesmittelwert.

  Insgesamt bewerteten 16 der 24 Befragten diesen Aspekt als positiv und nützlich. Daraus lässt sich schlussfolgern, dass der intendierte Nutzen der Lernstandserhebungen durchaus erkannt wird. Allerdings wurde dies nur in Ausnahmefällen tatsächlich für die eigene Unterrichtsentwicklung genutzt.

- Anregung zu verstärkter Kooperation

  SL 11: Also es bietet für die Schule das Potenzial zu Absprache, zu Austausch, zu Koordination […]. (LK 11 + SL 11, 00:54:46-4)

  Der Befragte erkannte den Mehrwert einer kooperativen Auswertung in Bezug auf den Informationsgewinn durch die Lernstandserhebungen. Allerdings konnte er in seiner Schule nicht feststellen, dass eine solche Zusammenarbeit vertieft stattgefunden hat.

- Anlass zur professionellen Selbstreflexion

  SL 5: Aber für mich war es, wie gesagt und so weiß ich das auch von den Kolleginnen, einfach so ein Feedback, wo die Schüler jetzt stehen und dass mit dem Unterricht, so wie man das macht, alles in Ordnung ist. (SL 5, 00:33:19-2)

Die Lernstandserhebung wurde als Anreiz wahrgenommen, eine Selbstreflexion vorzunehmen, inwiefern die Testergebnisse mit der eigenen Arbeit in Verbindung stehen. Dies regte die interne Attribuierung bei der Auswertung an.

- Anregungen durch neue Test- und Aufgabenformate

  LK 10: Gut, also bezogen auf das konkrete Beispiel Englisch in der achten Klasse, ein anderes Beispiel habe ich ja nicht, bieten die Lernstandskontrolle eine sehr gute Testmöglichkeit für standardisierte Prüfungssituationen, wie sie auch später im Landesabitur der Fall sein werden. (LK 10, 00:32:20-3)

  LK 2: Sie bringt weiterhin für mich so Anregungen im Aufgabenrepertoire, welche Möglichkeiten es gibt, bestimmte Schüler mit bestimmten Aufgaben zu konfrontieren und daraus Schlussfolgerungen zu ziehen. (LK 2, 00:55:56-0)

  Die Durchführung der Lernstandserhebung böte die Möglichkeit, die Lehrkräfte wie auch die Schüler mit einem externen Testformat zu konfrontieren. Dies sei insbesondere langfristig als Vorbereitung auf das Landesabitur eine wertvolle Erfahrung. Zugleich erhielten die Lehrpersonen Einblick in innovative Aufgabenstellungen, die kompetenzorientierten Anforderungen entsprechen. Die Testaufgaben und Aufgaben der didaktischen Hinweise könnten sie zukünftig als Unterrichtsmaterialien verwenden.

- Lernstandserhebungen als Bestandteil der Reformentwicklung

  SL 2: Aber sie sind ein Baustein in einer neuen Entwicklung, in der man Lernen von hinten, vom Ergebnis her stärker sieht. Das meine ich schon, da sind sie ein richtiger Weg, den man begehen soll. (SL 2, 00:53:01-9)

  Die Lernstandserhebungen wurden als ein in die Reformentwicklung der Standardisierung und Kompetenzorientierung integriertes Instrument zur Leistungsmessung verstanden. Der Schulleitungsvertreter 2 erhoffte sich eine konkretere Erläuterung des Grundkonzepts dieser Entwicklung für die Lehrkräfte mithilfe der Tests. Ein anderer Proband sah den Nutzen der Tests vor allem im Bildungsmonitoring, um Informationen zur Weiterentwicklung der Bildungsstandards zu generieren. Die Lernstandserhebungen werden in Hessen jedoch zu diesem Zweck nicht ausgewertet.

Die Probanden empfanden es bis auf eine Ausnahme als notwendig, dass die Lernstandserhebungen vor allem einen persönlichen Nutzen zur Folge haben sollten. Die Bewertung dieses Nutzens fiel jedoch heterogen aus. Während zwei Drittel der Befragten den diagnostischen Nutzen erkannten, führten einige von ihnen dennoch kritische Aspekte an, welche das Potenzial der Lernstandserhebungen massiv einschränken würden.

- Informationen und Anregungen sind bereits bekannt

  LK 8: Ja, ich hätte mir gewünscht, dass wir sagen: "Ach, ich habe jetzt etwas gelernt für meinen Unterricht und das kann ich in Zukunft weiter machen." Das haben diese Lernstandserhebungen nicht gebracht, fand ich. (LK 8, 00:28:40-4)

  Die Lernstandserhebungen boten für diesen Befragten keine neuen Erkenntnisse oder Anregungen. Somit konnte er trotz einer Auswertung auf mittlerem Intensitätsniveau in den Tests keinen persönlichen Nutzen für sich sehen. Als besonders problematisch wurde angeführt, dass die Übertragung der Informationen auf den Unterricht nicht möglich seien. Sinnvoller wäre hingegen eine individualisierte Gestaltung der Rückmeldungen.

- mangelhafte Ausrichtung des Unterricht auf die Kompetenzorientierung

  SL 5: Und ansonsten denke ich, wird es vielleicht ein bisschen interessanter, wenn man kompetenzorientiert unterrichtet, denke ich. So finde ich das momentan noch so ein bisschen abgekoppelt […]. (SL 5, 00:33:19-2)

  Da der alltägliche Unterricht bislang noch unzureichend auf die Kompetenzorientierung ausgerichtet sei, wurde eine Integration der Lernstandserhebungen in den Unterricht bislang noch als problematisch betrachtet. Dies hätte unter anderem den Nebeneffekt, dass einige getestete Kompetenzen noch nicht im Unterricht fokussiert werden. Aus dieser Aussage kann die Behauptung abgeleitet würden, dass mit zunehmender Kompetenzorientierung auch der Nutzen der Tests für die Unterrichtsentwicklung steigt.

- Ablehnung des Grundkonzepts einer standardisierten Leistungsmessung

  SL 9: Vergleichbarkeit, Operationalisierbarkeit. Das ist im Moment so das Gebot der Stunde. Ich halte das nicht unbedingt für einen Gewinn hinsichtlich des Lernzuwachses. (SL 9, 00:07:33-0)

  Der Befragte führte den vermeintlich kontraproduktiven Nutzen der Lernstandserhebungen an, indem durch die Konzentration auf den Leistungsvergleich und die Positionierung eine „Testeritis“ mit der Zielsetzung entstehe, möglichst gut abzuschneiden. Dies sei für die Nachhaltigkeit von Lernprozessen eher hinderlich.

Im Zusammenhang mit dem von den Probanden häufig angeführten Kritikpunkt des hohen Arbeitsaufwandes, der mit den Lernstandserhebungen verknüpft sei, wurde erfragt, inwiefern dieser Aufwand in Relation zu dem wahrgenommenen Nutzen steht. Hierbei konnten drei verschiedene Positionen bei den Befragten herausgefiltert werden.

Zum einen empfanden zwei Schulleitungsmitglieder den Aufwand definitiv als lohnend aufgrund des hohen Informationsgehalts der Rückmeldung, wie folgender Interviewausschnitt aufzeigt:

> *SL 1: […] Das war eigentlich eine relativ zumutbare Sache. Und was ich an Resultaten herauskriege, da lohnt sich die Arbeit allemal. Also ich glaube, hier ist die Zeit da gut investiert. (SL 1, 00:29:16-0)*

Der Zeitaufwand sei nicht höher als bei der Nutzung anderer kompetenzorientierter Diagnoseformen, wie der Auswertung von Selbstdiagnosebögen etc. Daher stand der Nutzen der Lernstandserhebung im Vordergrund der Wahrnehmung. Der Befragte hatte den Test selbst in Mathematik in der sechsten Klassenstufe durchgeführt und beurteilte die Korrektur als nicht besonders zeitaufwendig.

Zum anderen stand bei drei Befragten der Aufwand mit dem persönlichen Nutzen in etwa im Gleichgewicht. Das diagnostische Potenzial der Lernstandserhebungen, auch für die Schüler, wurde von ihnen erkannt. Dieser wöge den Aufwand auf, der zugleich als immens eingeschätzt wurde. Die Position dieser Personen ist somit eher ambivalent. Einerseits schätzten sie die aufgebrachte Zeitdauer als zu hoch ein; andererseits sahen sie einen Nutzen in dieser Arbeit, indem beispielweise Defizite der Klasse erkennbar wurden. Alle drei Befragten gehörten den Schulen an, die den Lehrkräften einen Korrekturtag bewilligt und somit die Mehrarbeit auf diese Weise honoriert bekommen haben.

Demgegenüber steht eine Gruppe von 12 Probanden gegenüber, für die der Arbeitsaufwand in keinem positiven Verhältnis zu dem Nutzen stand. Die nachfolgenden Zitate geben exemplarisch die Aussagen wieder:

> *LK 3: Ich konnte es für gar nichts gebrauchen. Und das war einfach ein Haufen Arbeit für nichts! (LK 3, 00:04:15-0)*

> *SL 8: Also der Korrekturaufwand, der steht dann schon nicht in so einem richtigen Verhältnis zu dem, was man dann für sich auch und für seinen Unterricht als Nutzen daraus ziehen kann. Deshalb hat die Fachkonferenz Deutsch auch beschlossen, im nächsten Jahr nicht mehr teilzunehmen. (SL 8, 00:08:11-7)*

Die Lehrkräfte bewerteten die Lernstandserhebungen dahingehend, dass die Arbeit umsonst hierein investiert worden war, da kein persönlicher Nutzen zu erkennen war. Sie verbanden mit den Tests hauptsächliche negative Assoziationen. Zudem hätte das Testinstrument eine zu geringe Bedeutung im Schulalltag, um intensiver die Ergebnisse auszuwerten. Diese Wahrnehmung führte in insgesamt fünf Schulen dazu, dass einige Lehrkräfte oder ganze Fachbereiche beschlossen, zukünftig nicht mehr an den Lernstandserhebungen teilzunehmen und sich somit diesen Mehraufwand zu ersparen.

Zuletzt wurde in den Interviews eine vergleichende Bewertung des Nutzens von den Lernstandserhebungen und den schulinternen Vergleichsarbeiten erfragt. Die thematische Ver-

knüpfung dieser beiden Leistungsmessungen liegt nahe, da beide Formen sowohl eine erhöhte Kooperation zwischen den Lehrkräften als auch einen Vergleich der Schülerleistungen für eine Stärken-Schwächen-Analyse anstreben. Die Befragten sollten sich in den Interviews nicht für eine der Leistungsmessungen entscheiden, sondern der empfundene Nutzen der Lernstandserhebung sollte ihnen durch diese Betrachtung bewusster werden. Bei dieser Diskussion wurde ein zentrales Argument angebracht, das einen höheren Nutzen bei den internen Parallelarbeiten aufgrund einer höheren Aussagekraft und spezifischeren Informationen zur Weiterarbeit impliziert.

LK 7: Was man bei den Vergleichsarbeiten nicht hat, ist diese Angst der Schüler davor, weil sie wissen, was auf sie zukommt, denn es ist eine Klassenarbeit und bei Klassenarbeiten kennen sie das Format und wissen, was da passiert. Und es ist natürlich auch so, dass die Vergleichsarbeiten auf die Schüler abgestimmt sind, das heißt, es ist dann nur Vokabular enthalten und es sind auch nur grammatische Strukturen enthalten, die die Schüler schon kennen. Das ist bei den Lernstandserhebungen eine Schwierigkeit […]. (LK 7, 00:31:17-0)

SL 5: […] Vergleichsarbeiten schreibt man ja nach einer Unterrichtseinheit, die man selbst konzipiert hat und die man durchgeführt hat und wo man einfach wissen will, […] wie ist jetzt der Lernstand der Schüler. Und ich finde es halt auch einfach umfassender. (SL 5, 00:20:11-3)

Die internen Vergleichsarbeiten böten nach Ansicht der Befragten mehr Aussagekraft und einen größeren Informationsgehalt. Ein Grund hierfür ist, dass die Inhalte der Arbeit konkret auf den eigenen Unterricht bezogen seien und daher keine unbekannten Aufgabenformate und thematischen Aspekte vorkämen. Somit würden die Parallelarbeiten für die Schüler und Lehrkräfte gewohnte Strukturen darstellen. Zum anderen würden die Fähigkeiten vertieft zu einer Unterrichtseinheit abgefragt. Demnach seien die Ergebnisse umfassender und spezifischer, wodurch gezieltere Hinweise auf Defizite und für die Weiterarbeit gewonnen werden könnten. Die Lernstandserhebungen seien hingegen weniger detailliert, da sie verschiedene Kompetenzbereiche nur oberflächlich abtesten würden.

Diese Argumente brachten insgesamt 13 Befragte vor. Dies deutet darauf hin, dass ihnen die eigentliche diagnostische Funktion der Lernstandserhebungen klar ist, sie diese im Vergleich zu dem Ertrag aus anderen Formen der Leistungsmessung jedoch als weniger aussagekräftig empfanden.

Gleichermaßen wurden Aspekte von den Befragten angeführt, die den besonderen Nutzen der Lernstandserhebung gegenüber den internen Vergleichsarbeiten herausstellten:

SL 3: Na die Vergleichsarbeit ist halt nur eine Klassenarbeit von vielen, die halt nur einen kleinen Teilbereich abdeckt. Und die Lernstandserhebung deckt einen ganz großen Teil ab, zwar auf einem anderem Niveau als die Klassenarbeit, aber jetzt was meine Klasse wirklich kann, denke ich, kriege ich bei den Lern-

standserhebungen mehr mit als bei einer Klassenarbeit, die sich über sechs Wochen erstreckt. (SL 3, 00:34:23-5)

Es wurden die inhaltliche Breite der Lernstandserhebungen sowie die Überprüfung von langfristig gesicherten und vernetzten Kompetenzen als positiv bewertet. Daher sei die Auswertung der Testergebnisse diagnostisch interessanter. Dies stellt eine konträre Position zu dem zuvor angebrachten Argument dar, was wiederum die Vielfältigkeit der Bewertung zu den Lernstandserhebungen durch die Befragten betont.

Des Weiteren wurde positiv wertend angeführt, dass keine Vorbereitung auf die Lernstandserhebung stattfände, währenddessen für die Vergleichsarbeit konkret im Unterricht geübt würde. Die Ergebnisse der Lernstandserhebungen seien demnach unabhängiger und aussagekräftiger, da sie einen nachhaltigen Kompetenzerwerb wiederspiegeln und aufzeigen würden, wie die Schüler auf neuartige Anforderungssituationen reagieren. Der Vergleich mit dem hessischen Landeswert wurde ebenfalls als ein Mehrwert gegenüber den Parallelarbeiten von zwei Befragten herausgestellt. Zudem führten zwei weitere Gesprächspartner an, dass sie durch die Lernstandserhebungen mehr innovative Anregungen in Hinblick auf Aufgabenstellungen etc. erhalten hätten, wie folgender Interviewausschnitt resümierend darlegt:

LK 2: Also die Anregungen, die ich von den Aufgabenformaten bekommen habe, gehen weit über das hinaus, was von der Vergleichsarbeit kam. (LK 2, 00:42:06-5)

12 Typisierung der Nutzungsformen der Probanden

Nachdem die Nutzungsprozesse, deren Einflussfaktoren sowie die Bewertung der Lernstandserhebungen durch die Probanden im Rahmen der qualitativen Inhaltsanalyse ausgewertet wurden, werden abschließend die Probanden den Nutzungsformen zugeordnet.

Es erfolgt zunächst ein Rückgriff auf das Modell der Nutzungsformen von Rossi, et al. (vgl. 2004, S. 411 f.). Hierbei wird zwischen der symbolischen, konzeptionellen und instrumentellen Nutzung differenziert (vgl. Abschnitt 5.5). Bei der *symbolischen* Nutzungsform wird die Rückmeldung zu den Testergebnissen hauptsächlich selektiv verwendet, um eigene Einschätzungen und Beurteilungen zu bestätigen. Eine Handlungsbereitschaft wird dabei nicht generiert. Dies bedeutet, dass die Aktionsphase nur ein minimales Ausmaß annimmt. Aussagen über die Intensität der Rezeption und Reflexion können damit jedoch nicht abgeleitet werden. Die *konzeptionelle* Nutzung beeinflusst hingegen das Denken oder die Einstellungen des Probanden. Die Aktion wird hierbei weniger durch konkrete Handlungen sichtbar, sondern vor allem unbewusste Wirkungen sind das Resultat des Nutzungsprozesses. Bei der *instrumentellen* Nutzungsform werden konkrete Handlungen aus den Ergebnissen abgeleitet und durchgeführt.

Die Nutzungsweise der Probanden der Interviewstudie konnte jeweils diesen drei Formen zugewiesen werden. Teilweise lagen Mischformen vor, indem zwei Nutzungsformen Verwendung fanden. In diesen Fällen wurden die Befragten derjenigen Nutzungsform zugeordnet, welche primär Verwendung fand. Der nachfolgenden Tabelle 14 kann die Zuweisung der Probanden zu den Nutzungsformen entnommen werden.

Zuordnung der Probanden zu den Nutzungsformen nach Rossi, et al.		
Nutzungsform	**Charakterisierung**	**Zugeordnete Personen**
symbolische Nutzung	selektive Nutzung zur Bestätigung der eigenen Einschätzung; kaum Aktion	LK 1, LK 3, LK 7, LK 8, LK 11, LK 12; SL 2, SL 4, SL 6, SL 7, SL 8, SL 9, SL 11
konzeptionelle Nutzung	unbewusste Wirkungen auf Einstellungen der Person	LK 4, LK 6; SL 3, SL 5, SL 10
instrumentelle Nutzung	Ableitung und Umsetzung von Maßnahmen	LK 2, LK 10, LK 9, LK 5; SL 1, SL 12

Tabelle 14: Zuordnung der Probanden zu den Nutzungsformen nach Rossi, et al. (vgl. 2004, S. 411 f.)

Demnach ergriff über die Hälfte der Befragten eine symbolische Nutzungsform für den Umgang mit den Lernstandserhebungen. Dies ist insofern von Bedeutung, da erneut die mangelnde Umsetzung der Aktionsphase sichtbar wird. Lediglich bei sechs Personen konnte eine instrumentelle Nutzung, welche sich in konkreten Handlungen zeigt, festgestellt werden. Selbstverständlich sind auch die symbolische und konzeptionelle Nutzungsform relevant, denn hierbei werden Ergebnisse ausgewertet und gedeutet. Zudem kann insbesondere die Professionalisierung bei der konzeptionellen Nutzung gefördert werden. Eine nach außen sichtbare Weiterentwicklung der Schulqualität und insbesondere der Unterrichtsqualität wird jedoch ausschließlich bei der instrumentellen Nutzung erreicht.

Ein Zusammenhang zwischen der Nutzungsform und dem jeweiligen Unterrichtsfach, in dem die Lernstandserhebung durchgeführt wurde, konnte nicht festgestellt werden. Dies verstärkt das Ergebnis, dass die individuelle Nutzungsintensität insbesondere von der Professionalität und der intrinsischen Motivation der Lehrkraft abhängig ist und weniger von dem fachspezifischen Testheft und dessen Rückmeldung.

Von den zwölf befragten Schulleitungsvertretern führten lediglich sieben die Lernstandserhebung selbst in einer ihrer Klassen durch. Die übrigen Schulleitungsmitglieder haben daher noch keine persönlichen Erfahrungen in der Durchführung und Korrektur der Tests gewinnen können. Dennoch können sie zum Beispiel die Ergebnisse der Klassen innerhalb der Schulleitung oder mit einzelnen beteiligten Lehrpersonen ausgewertet und Maßnahmen zur kooperativen Auswertung der Schülerresultate im Kollegium angeregt haben. Es ist auffällig, dass vier von diesen fünf Schulleitungsvertretern eine symbolische Nutzung vorgenommen haben und somit keine weiteren Handlungen initiiert haben.

Bei der instrumentellen Nutzung weisen drei der dort vertretenen Lehrpersonen eine besondere akademische Qualifikation auf (z.B. Fachausbilder am Studienseminar, Didaktiker an Universitäten etc.). Dies verstärkt erneut die Erkenntnis, dass die Nutzungsintensität und -qualität von dem Professionsgrad der agierenden Person abhängig ist.

Neben den Nutzungsformen nach Rossi, et al. (vgl. 2004, S. 411 f.) wurde im Abschnitt 5.6 eine weitere Typisierung von Hartung-Beck (vgl. 2009, S. 125 ff.) dargelegt. Dabei wurde zwischen vier Professionstypen sowie vier Organisationstypen in Bezug auf den Umgang mit den Lernstandserhebungen differenziert. Es wurde in dieser Untersuchung der Versuch unternommen, die Probanden diesen Professions- und Organisationstypen ebenfalls zuzuweisen. Dies war allerdings nicht möglich, da die Typenbeschreibungen zu speziell waren und die Aussagen der Interviewpartner ihnen nicht eindeutig zugeordnet werden konnten.

In einer letzten Auswertung bezüglich der Nutzungsformen bei den Probanden erfolgt ein Rückgriff auf die Funktionen der Lernstandserhebungen. Bei einer Fokussierung auf die

zentrale Aufgabe, der Anregung von Unterrichtsentwicklung, kann zwischen summativen und formativen Funktionen differenziert werden (vgl. Abschnitte 4.2 und 5.4). Während bei der summativen Nutzung die Lernstandserhebung vorwiegend als ein Diagnoseinstrument verwendet wird, erfolgt bei der formativen Nutzung eine konkrete Förderung der Unterrichtsentwicklung. Bei einer Zuordnung der Probanden hinsichtlich ihrer Nutzung der Lernstandserhebung zu diesen zwei Kategorien ergibt sich folgende Aufteilung (vgl. Tabelle 15):

Zuordnung der Probanden zu summativer und formativer Nutzung der Lernstandserhebung	
summative Nutzung als Diagnoseinstrument	LK 3, LK 4, LK 6, LK 7, LK 8, LK 9, LK 10, LK 11, LK 12; SL 2, SL 3, SL 4, SL 5, SL 6, SL 7, SL 8, SL 9, SL 11, SL 12
formative Nutzung zur Unterrichtungsentwicklung	LK 1, LK 2, LK 5; SL 1, SL 10

Tabelle 15: Zuordnung der Probanden zu summativer und formativer Nutzung der Lernstandserhebung

Während lediglich fünf Befragte eine formative Nutzung aufwiesen, wurden die Rezeptions-, Reflexions- und Aktionsphasen bei 19 Personen der summativen Nutzung zugewiesen. Dies drückt sehr deutlich aus, wie der Zweck der Lernstandserhebungen bei den Lehrkräften und Schulleitungsvertretern vorwiegend wahrgenommen wurde.

TEIL D - FAZIT

13 Zusammenfassung der Untersuchungsergebnisse

In der vorgenommenen qualitativen Untersuchung zur Nutzung der Lernstandserhebungen an sechs hessischen Gymnasien wurden zwölf Lehrkräfte sowie zwölf Mitglieder der erweiterten Schulleitung in Form von leitfadengestützten Interviews befragt. Die Auswertung wurde anhand der qualitativen Inhaltsanalyse durchgeführt und die Ergebnisse in den Abschnitten 8 bis 12 ausführlich dargelegt. Bevor die zugrundeliegenden Fragestellungen beantwortet werden, erfolgt eine Zusammenfassung der wesentlichen Erkenntnisse.

Die Vergleichsarbeiten in Hessen, welche in diesem Bundesland als „Lernstandserhebungen“ bezeichnet werden, wurden mit dem Schuljahr 2006/2007 schrittweise implementiert. Zeitlich betrachtet wurden die Schulen mit den Lernstandserhebungen vor der Einführung der Bildungsstandards konfrontiert. Die Teilnahme in der Sekundarstufe I beruht überwiegend auf Freiwilligkeit. Die wesentliche Kommunikation zwischen der Einzelschule und dem zuständigen Institut (LSA) in Wiesbaden erfolgt über einen Schulkoordinator. Zusätzlich sind die teilnehmenden Lehrkräfte gebeten, nach Erhalt der Rückmeldung einen Fragebogen online auszufüllen. Die Schülerergebnisse werden nicht veröffentlicht und anonymisiert ausgewertet. Die Lehrpersonen erhalten zur Unterstützung des Nutzungsprozesses didaktische Zusatzmaterialien sowie die Aufbereitung der Schülerleistungen in Form einer Rückmeldung. Diese enthält eine tabellarische Auswertung, einen Sofortbericht sowie einen Ergebnisbericht, welcher hessische Vergleichswerte enthält. Es erfolgt jedoch keine Bewertung der Ergebnisse in Form einer kriterialen Zuordnung zu Kompetenzstufen. Vielmehr werden prozentuale Durchschnittsergebnisse mittels Prozentwerten angegeben, die keine Individualdiagnostik implizieren. Knappe und standardisierte Interpretationshinweise sind den Grafiken der Rückmeldung beigefügt.

Rezeptionsphase

Entsprechend dem Zyklenmodell nach Helmke (Helmke A. , 2004) wurden die Aussagen der Interviewprobanden entsprechend den einzelnen Nutzungsphasen ausgewertet. In der *Rezeption* wurden die Testinhalte hinsichtlich der angesprochenen Kompetenzbereiche, der Aufgabenschwierigkeiten, der Lehrplanadäquatheit und der Aufgabenformate von den Gesprächspartnern betrachtet. Daher setzte die Rezeption teilweise bereits vor dem Erhalt der Rückmeldung ein. Die Rezeption erfolgte zu verschiedenen Zeitpunkten und ist aus diesem Grund nicht als ein zeitlich zusammenhängender Prozess zu verstehen.

Die zeitlich vorgelagerte Rezeption beinhaltete unter anderem eine tiefgründige Betrachtung in Form einer aufgabenbezogenen Auswertung. Im Zentrum des Interesses stand jedoch die Rückmeldung mit den Schülerergebnissen. Der Sofortbericht war bei einigen Lehr-

kräften weniger von Bedeutung, was sich negativ auf die Intensität der Auseinandersetzung mit dem Ergebnisbericht auswirkte. Ein Großteil der Lehrkräfte fokussierte sich in der Rezeption primär auf den sozialen Vergleich des Gesamtergebnisses mit dem hessischen Vergleichswert. Dies stellte die interessanteste Information der Rückmeldung dar. Insbesondere bei den Vertretern der Schulleitungen war eine hohe Bedeutsamkeit des sozialen Vergleichs festzustellen. Lediglich drei Interviewpartner interpretierten die Rückmeldung als eine Grundlage für eine individuelle Schülerauswertung.

Des Weiteren war festzustellen, dass nicht alle beteiligten Lehrkräfte bedingt durch verschiedene Gründe eine Rückmeldung erhielten. Dennoch erfolgte in einem Fall eine intensive Reflexion. Daher scheint die Rezeption der Rückmeldung für eine tiefgründige Reflexion nicht zwingend erforderlich zu sein.

Die Rezeption der didaktischen Materialien erfolgte in wenigen Fällen nur sporadisch, so dass keine Wirkungen auf die Unterrichtsentwicklung festzustellen waren. Bezüglich der Intensität der Rezeption wurden in der Untersuchung immense Unterschiede konnotiert. Als zentrale hemmende Einflussfaktoren konnten Zeitmangel, eine fehlende intrinsische Motivation, eine eher ablehnende Bewertung der Testkonzeption sowie negative Erfahrungen im Kollegium ermittelt werden. Innerhalb der Teilnehmergruppe der Untersuchung gab es niemanden, der keine Form der Rezeption durchgeführt hätte. Jedoch kann dies nicht als allgemeingültig interpretiert werden, da Aussagen von Schulleitungsmitgliedern darauf hindeuten, dass durchaus einige Lehrkräfte die Lernstandserhebungen mitsamt ihrer Rückmeldungen und Materialien vollständig ignorierten.

Reflexionsphase

Im Rahmen der *Reflexion* beobachteten einige Probanden einen Zusammenhang zwischen dem Schülerverhalten und den Testergebnissen, indem sich Konzentration, Angst, etc. auf die Leistung ausgewirkt hätten. Der zentrale Indikator für die Reflexion stellte die persönliche empfundene Zufriedenheit mit den Schülerleistungen dar. Eine positive Zufriedenheit wurde als Bestätigung der eigenen pädagogischen und fachlichen Arbeit interpretiert, so dass die Selbstreflexion angeregt wurde. Für das Maß der Zufriedenheit diente überwiegend der soziale Vergleich mit dem hessischen Referenzwert als zentrale Bezugsnorm. Dies erwies sich allerdings als ein eher oberflächlicher Vergleich der Ergebnisse ohne eine Verknüpfung mit der inhaltlichen Bewertung der Schülerleistungen. Einige Lehrpersonen nahmen Stärken-Schwächen-Analysen vor, denen erneut der soziale Vergleich als Grundlage diente. Das Erkennen von Stärken implizierte wiederum eine positive Bestätigung, löste jedoch keinen Handlungsimpuls aus. Bei defizitären Leistungen wurde stets eine intensivere

Ursachenanalyse vorgenommen, wobei diese seltener mit der eigenen Arbeit begründet wurde.

Bei etwa der Hälfte der Gesprächspartner wurden in diagnostischer Sicht die eigenen Einschätzungen zu den Schülerleistungen grob bestätigt. Insbesondere der Vergleich zu den bisherigen Bewertungen der Schüler beeinflusste den Erkenntnisgewinn durch die Lernstandserhebung in positiver Weise maßgeblich. Dies stellte eine Chance dar, neue Perspektiven zu den Schülerleistungen zu erhalten, was besonders bei einer Abweichung zu den bisherigen Eischätzungen zu einer intensiven Reflexion führte. Als Gründe für solche Abweichungen wurden die Form die Aufgabenstellungen, die Durchführungs- und Auswertungsrichtlinien des Tests sowie die Vernachlässigung bestimmter Kompetenzbereiche im eigenen Unterricht angeführt.

Vergleiche mit anderen Leistungsdaten (Parallelklassen, Vergleichsarbeiten etc.) wurden selten und ausschließlich von Schulleitungsmitgliedern beziehungsweise Lehrkräften in der Funktion eines Fachsprechers vorgenommen. Insgesamt konnten sieben von 19 Gesprächspartnern, die an den Lernstandserhebungen beteiligt waren, keine neuen Erkenntnisse erlangen, so dass die eigenen Einschätzungen bestätigt wurden. Dies hatte zur Folge, dass kein Handlungsimpuls im weiteren Nutzungsprozess entstand. Die in den Rückmeldungen enthaltenen Anregungen für die Weiterarbeit konnten ebenfalls nicht als wirksame Hilfe dienen, da sie als nicht individuell genug bewertet wurden.

Bezüglich der Attribuierung der Schülerergebnisse ist festzustellen, dass positive Ergebnisse fast ausschließlich internal attribuiert wurden, indem sich die Testschwerpunkte auch im Unterricht wiederfinden würden. Bei defizitären Ergebnissen ergab sich ein gemischteres Bild: Sie wurden internal attribuiert, wenn eine Diskrepanz zwischen den pädagogischen Prinzipien der Leistungsmessung und dem eigenen professionellen Verständnis festgestellt wurde beziehungsweise wenn bestimmte Themen- und Kompetenzbereiche im Unterricht vernachlässigt wurden. Zusätzlich wurden externale Begründungen, wie die Klassenzusammensetzung, die Motivation und Konzentration der Schüler sowie die Testkonzeption im Allgemeinen (Inhalte, Zeitpunkt, Korrekturbestimmungen etc.) angeführt. Falls die Lernstandserhebung eher ablehnend bewertet wurde, wurden auch die Ergebnisse der Lernenden external begründet. Eine internale Ursachenreflexion wurde auf diese Weise vermieden, was eine Abgabe der Verantwortung für die Schülerleistungen implizierte.

Die Reflexionsintensität war bei den Schulleitungsvertretern geringer als bei den Lehrkräften. Der Grund hierfür ist die bereits erwähnte Fokussierung auf die Positionierung der Klasse im Vergleich zum hessischen Referenzwert. Bei Zufriedenheit bestand kein weiteres Erkenntnisinteresse. Eine Attribuierung erfolgte vor allem bei defizitären Ergebnissen. Bis

auf eine Ausnahme konnte bei allen Befragten eine Reflexion festgestellt werden. Dabei war eine geringe Rezeption nicht primär ausschlaggebend für die Intensität der Reflexion. Allerdings konnte durchaus beobachtet werden, dass eine intensive Rezeption wiederum eine intensive Reflexion beförderte. Falls die Testkonzeption als negativ bewertet wurde, hatte dies eine Einschränkung der Reflexion zur Folge. Als Gründe für eine eher oberflächliche Reflexion wurden zudem mangelnde Zeit und Motivation sowie der geringe Stellenwert der Lernstandserhebung im Unterrichtsalltag angeführt.

In Bezug auf eine Kommunikation der Ergebnisse während der Reflexion mit weiteren schulischen Akteuren konnte bei den Befragten generell das Bedürfnis nach einem Austausch mit Kollegen über die Durchführung, die Korrektur und die persönlichen Erfahrungen festgestellt werden. Nach Ansicht der Lehrkräfte sollte den Lernstandserhebungen in diesem Bereich eine größere Bedeutung beigemessen werden. Durch eine intensivere kommunikative Zusammenarbeit könnte auch der Nutzen der Leistungsmessung für die unterrichtliche Arbeit erhöht werden.

Allerdings wurden beispielsweise Vergleiche mit Parallelklassen nur in Einzelfällen durchgeführt. Auch die Testkonzeption war ein geringer Gesprächsanlass. Die in der Regel inoffiziellen Gespräche fanden ausschließlich bei negativen Erfahrungen statt. Als hindernde Faktoren für eine gezieltere Kommunikation wurden der Zeitpunkt der Lernstandserhebungen im Schuljahr sowie die geringe Anzahl von teilnehmenden Kollegen desselben Faches angeführt. Zudem hat die vereinzelt stattgefundene Vergabe der Korrektur an externe Kräfte die Kommunikationsintensität negativ beeinflusst, da in diesen Fällen keine Gesprächsgrundlage vorhanden war.

In den Fachkonferenzen waren die Schülerergebnisse oftmals Gegenstand der Besprechungen, wobei dies jeweils eine eher oberflächliche Diskussion über die Positionierung im hessischen Vergleich implizierte. Vereinzelt wurde in den Konferenzen über die weitere Teilnahme an den nächsten Testdurchläufen abgestimmt. In Gesamtkonferenzen stellten die Lernstandserhebungen fast nie ein Thema dar. Lediglich in einem Fall wurde die Testkonzeption auf der Gesamtkonferenz erläutert. Konstruktive Diskussionen fanden daher oft nicht statt. Auch eine intensive Reflexion hatte keine intensive Kommunikation mit anderen schulischen Akteuren zur Folge, da teilweise die Kooperationsstrukturen hierzu fehlten.

Bei der Hälfte der untersuchten Schulen fand eine Besprechung auf Schulleitungsebene statt. Diese Kommunikationsform besaß stets Berichtscharakter zu der Testkonzeption, verknüpft mit einer oberflächlichen Betrachtung der Gesamtergebnisse. Bei der anderen Hälfte der teilnehmenden Schulen waren Desinteresse der Schulleitung zu dieser Thematik

sowie der geringe Stellenwert der Lernstandserhebung die Gründe für die fehlende Kommunikation.

Mit den Schülern, deren Leistungen getestet wurden, fand selten eine individuelle Auswertung statt. Die Konzeption der Rückmeldung mitsamt ihrer Aufbereitung der Ergebnisse ließ dies nicht zu. Somit konnten die Schüler ihre eigene Leistung kaum bis gar nicht verorten. Nur vier Befragte führten an, persönliche Gespräche mit einzelnen Lernenden geführt zu haben. Vielmehr waren die Einschätzungen und Erfahrungen der Schüler während der Testdurchführung Gesprächsanlässe im Unterricht. Diese umfassten maximal eine Unterrichtsstunde. Dabei wirkten sich weniger gute Assoziationen der Lernenden auf die Grundhaltung der Lehrkraft zu den Tests negativ aus.

Ein Großteil der befragten Lehrpersonen kommunizierte nicht mit den Schülern über die Lernstandserhebung. Die große Zeitspanne zwischen dem Testzeitpunkt und dem Erhalt der Rückmeldung war ein häufig angeführter Grund hierfür. Des Weiteren wurden die Bewertungsrichtlinien als für die Schüler irreführend angesehen, so dass eine Besprechung der Testleistungen keinen effektiven Nutzen für die Schüler bringen würde.

Des Weiteren ist zu erwähnen, dass in zwei Schulen die Rechenschaftsfunktion der Lernstandserhebungen verstärkt herausgestellt wurde, indem Zeitungsartikel über das Abschneiden der Schüler an dem Test erschienen.

Den Online-Fragebogen zur Rückmeldung an das zuständige Institut LSA nutzten nur acht Befragte als Kommunikationsinstrument. In diesem Zusammenhang wurde immense Kritik geäußert, da der Fragebogen auf den Nutzungsprozess der Lernstandserhebungen, nicht aber auf die Konzeption und Durchführung der Tests ausgerichtet ist. Dieser Feedbackkanal wurde mehrheitlich ablehnend in seiner Qualität und Konzeption beurteilt.

Aktionsphase

Im Zyklenmodell schließt sich an die Reflexion der Ergebnisse die *Aktion* an, in der konkrete Handlungen im Rahmen der Schulentwicklungen eingeleitet werden. Bezogen auf die Unterrichtsentwicklung fand bei einigen Probanden eine Bewertung der Testergebnisse statt, indem eine Teilnote erstellt wurde. Auf diese Weise sollte die Arbeit der Schüler sowie der Lehrkräfte honoriert und die Schüler zusätzlich motiviert werden. Das Problem liegt hierbei in der eigenhändigen Festlegung des Bewertungsmaßstabes. Letztlich ist diese Handlung eine selektive Auswertung der Lernstandserhebung, deren Aussagekraft auf die Note reduziert wird.

Acht beteiligte Gesprächspartner erhielten zudem Anregungen durch die Aufgabenformate. Diese Hinweise wurden jedoch nicht praktisch umgesetzt, da geschlossene Testitems zu-

nächst in offene Aufgabenstellungen transformiert werden müssten. Ein einem Fall wurde von einem gezielten Einsatz der Aufgabenformate im Unterricht im Sinne eines „Teaching to the Test“ von einem Schulleitungsvertreter berichtet. Teilweise wurden auch förderdiagnostische Potenziale in den Testergebnissen erkannt, die allerdings nur von einem Befragten umgesetzt wurden. Hemmend wirkte hierauf die Aufbereitung der Rückmeldung, welche primär Klassenwerte und nicht Individualwerte beinhaltet. Auch das Bewusstsein, dass bestimmte Kompetenzbereiche wiederholt oder stärker im Unterricht berücksichtigt werden müssten, führte bei lediglich einem Gesprächspartner zu einer entsprechenden Handlung.

Insgesamt berichtete zwar etwa die Hälfte der Probanden von inhaltlichen Anregungen, jedoch konnte nur bei vier Personen Maßnahmen dahingehend festgestellt werden. Als Gründe für die mangelnde Umsetzung wurden die fehlende Zeit, der Zeitpunkt der Tests sowie die Zufriedenheit mit den Ergebnissen, welche keinen Handlungsbedarf erfordern würden, angeführt. Bei einem Großteil der Befragten fand somit keine direkte Aktion statt. Demgegenüber konnten nur wenige punktuelle Maßnahmen konnotiert werden, die oft keine Unterrichtsentwicklung beförderten.

Eine intensive Reflexion hat demnach unmittelbar nicht eine mittlere bis hohe Aktion zur Folge. Lediglich bei fünf Gesprächspartnern war eine konstante Nutzungsintensität in den einzelnen Phasen zu erkennen. Demgegenüber wiesen zwölf Probanden eine abfallende Nutzungsintensität auf. Der Übergang von der Reflexion zur Aktion ist somit eine immense Herausforderung für eine wirkungsvolle und nachhaltige Nutzung der Lernstandserhebungen.

Im Kontext der Organisationsentwicklung wurden vorrangig vereinzelte Maßnahmen zur Verbesserung der organisatorischen Abwicklung des nächsten Durchlaufs der Lernstandserhebungen ergriffen, indem beispielsweise der Korrekturaufwand durch Korrekturtage oder eine externe Korrektur reduziert werden sollte. Kooperationsstrukturen für eine gemeinsame Auswertung der Tests wurden in keiner Schule ausgebaut oder initiiert. Gründe hierfür waren der geringe Stellenwert der Tests, die niedrige Anzahl der teilnehmenden Kollegen sowie die nicht ausreichend zur Verfügung stehende Arbeitszeit.

Bezüglich einer Personalentwicklung haben einige Lehrkräfte durch die Lernstandserhebung einen Fortbildungsbedarf bei sich diagnostiziert, jedoch dazu keine konkreten Maßnahmen ergriffen. Generell fand durch die Anregungen und die Umsetzung von Handlungen eine Weiterentwicklung der Lehrerprofessionalität in unterschiedlichem Ausmaß statt.

Evaluation

Die abschließende Phase der Evaluation konnte lediglich bei einem Probanden konnotiert werden, welcher aufgrund defizitärer Leistungen bei einem Item eine ähnliche Aufgabe in einer Klassenarbeit stellte und die Leistungen miteinander verglich. Ansonsten fand keine Evaluation statt und die Handlungen blieben unreflektiert.

Einflussnehmende Bedingungen

Neben den verschiedenen Nutzungsphasen wurden in der Untersuchung Einflussfaktoren untersucht, welche auf den Handlungsprozess der Befragten eingewirkten. Hierzu zählen zum einen die individuellen Bedingungen. Signifikant deutlich wurde in diesem Zusammenhang die Bedeutung der intrinsischen beziehungsweise extrinsischen Motivation zur Teilnahme an den Lernstandserhebungen. Da die Leistungsmessung in Hessen vorrangig auf Freiwilligkeit beruht, sind die Gründe für die Teilnahme wichtige Einflussfaktoren. Die Schulleitungsmitglieder wiesen erwartungsgemäß stets eine intrinsische Motivation auf, da sie der Teilnahme an den Tests zustimmen müssen. Bei den Lehrkräften konnte jedoch kein eindeutiges Bild erkannt werden. Einige Probanden wiesen eine intrinsische Motivation auf, während andere zum Beispiel aufgrund der Festlegung der Teilnahme durch die Schulleitung extrinsisch motiviert waren.

Als Gründe für die intrinsisch motivierte Teilnahme wurden das Bedürfnis nach einer Positionierung der Schülerleistungen, die zunehmende zukünftige Bedeutung der Bildungsstandards, die Hoffnung nach einer Weiterentwicklung der eigenen Lehrerprofessionalität sowie die Passung des Tests mit dem Förderkonzept der eigenen Schule angeführt. Demgegenüber können als Einflussfaktoren für eine extrinsische Motivation die Anzahl der teilnehmenden Lehrkräfte im Kollegium, die bisherigen Erfahrungen von Kollegen sowie die Einflussnahme der Schulleitung auf die Entscheidung zur Teilnahme im Sinne einer „Scheinfreiwilligkeit“ angeführt werden. Die Ausprägung der intrinsischen beziehungsweise extrinsischen Motivation hat Einfluss auf die innere Arbeitshaltung und die Intensität des Nutzungsprozesses bei den Lernstandserhebungen.

Als weitere individuelle Bedingung ist die individuelle Innovationsbereitschaft zu nennen, welche die Nutzung sowohl positiv als auch negativ beeinflussen kann. Insbesondere in letzterem Fall entstanden bei einigen Probanden aufgrund einer mangelnden Innovationsbereitschaft Zweifel am Mehrwert der Lernstandserhebungen. Allerdings hatte die Bewertung des Potenzials der Lernstandserhebung nicht immer direkte Auswirkungen auf die Nutzungsintensität, da die endgültige Bewertung in der Regel erst nach Abschluss der

Nutzung von den Lehrkräften vorgenommen wurde und damit eine Bestätigung oder Widerlegung der vorherigen Annahmen und Erwartungen erfolgte.

Als eine positive individuelle Bedingung wirkte sich bei den Befragten eine zusätzliche Qualifikation in der Professionalität aus, wie die Tätigkeit als Didaktiker, Fachausbilder, etc. Bei diesen Personen wurde jeweils eine intensive Reflexion festgestellt und auch die Aktionsphase zeichnete sich bei ihnen durch konkrete Handlungen aus. Demnach führen zuvor erworbene Kenntnisse zur Kompetenzorientierung im Unterricht zu einer intensiveren Nutzung.

Zum anderen wurden in der Auswertung der Interviews schulische Bedingungen betrachtet. Dabei konnte festgestellt werden, dass bei den zwölf untersuchten Gymnasien geringe bis keine Kooperationsstrukturen vorlagen, die für eine kooperative Auswertung der Lernstandserhebungen hätten genutzt werden können. Jedoch auch vorhandene Strukturen wurden von den betreffenden Lehrpersonen äußerst geringfügig genutzt. Lediglich schulische Zugeständnisse im Sinne einer Aufwandsentschädigung durch die Schulleitung, wie Korrekturtage, hatten eine positive Wirkung auf die Intensität aller Nutzungsphasen.

Als einflussnehmende externe Bedingungen sind die Freiwilligkeit sowie der Zeitpunkt der Lernstandserhebungen anzuführen, was von den Probanden kontrovers diskutiert wurde. Auch die Festlegung der zu testenden Klassenstufen beeinflusste die Nutzung, indem zum Beispiel die Dopplung externer Leistungsüberprüfungen in einer Klassenstufe negativ bewertet wurde. Diese Aspekte wirkten sich entsprechend positiv beziehungsweise negativ auf die Haltung des Probanden dem Test gegenüber aus.

Während der Interviews gingen die Befragten stets von sich aus auf ihre persönliche Bewertung der Konzeption der Lernstandserhebungen ein, so dass dieser Aspekt ebenfalls in die qualitative Auswertung der Untersuchung aufgenommen wurde. Die Rahmenbedingungen zur Durchführung und Korrektur der Tests stellten dabei einen Gesichtspunkt dar. Während individuell aufgetretene Probleme, zum Beispiel bei der Ergebniseingabe, lediglich Einzelfälle darstellten, wurden der Umfang und die zur Verfügung stehende Zeit für die Bearbeitung der Leistungsmessung von einer großen Anzahl der Probanden als unpädagogisch empfunden. Dies führte dazu, dass auch die zugrunde liegenden Zielsetzungen der Lernstandserhebung teilweise infrage gestellt wurden. Ein weiterer Aspekt, welcher angesprochen wurde, war die Aufgabenschwierigkeit sowie der inhaltliche Gegenstand der Tests. Als mehrheitlich positiv wurden in diesem Zusammenhang die Lehrplanadäquatheit, die Breite der getesteten Inhalte sowie die Aufgabenkontexte bewertet. Demgegenüber wurde kritisiert, dass zentrale Themen des Unterrichts keinen Raum in der Lernstandserhebung finden würden, was den Wert des Tests reduziere. Auch wurden das Vorhandensein von Gütekriterien für

die Items sowie die Aufgabenformulierungen ablehnend beurteilt. Die Testschwierigkeit wurde nicht eindeutig eingeschätzt, jedoch überwog eine Beurteilung als adäquat. Eine negative Bewertung hatte dabei Einfluss auf die Akzeptanz dieses Leistungsmessungsinstruments. Wenn der Schwierigkeitsgrad nicht den eigenen Erwartungen entsprach, wurde der Test als nicht geeignet beurteilt, was teilweise Auswirkungen auf die Entscheidung zur weiteren Teilnahme hatte. Letztlich konnte die Tendenz ermittelt werden, dass Items eher als zu leicht empfunden wurden, wenn die Kompetenzdiagnostik im Interesse der Lehrkraft bei der Nutzung stand. War ihr jedoch eine Positionierung der Schülerleistungen wichtig, wurden die Aufgaben tendenziell als zu schwer eingeschätzt.

Auch die Korrekturbestimmungen beeinflussten die Gesamtbewertung der Lernstandserhebungen. Es konnten Probleme und Unsicherheiten bei der Korrektur offener Items beobachtet werden, die zudem mehr Arbeitszeit in Anspruch genommen hatten. Folglich wurde die Aussagekraft der Tests angezweifelt, da eine Objektivität nicht gewährleistet sei und die Ergebnisse verfälscht seien. Letzteres geschah in einem Fall sogar wissentlich. Auch wurden einige Korrekturbestimmungen, wie die Einteilung in Richtig oder Falsch sowie die Nichtbeachtung der Rechtschreibung, als unpädagogisch und im Widerspruch zur üblichen Bewertungspraxis stehend empfunden. Die aufgewandte Korrekturzeit wurde mehrheitlich als unverhältnismäßig hoch eingestuft.

Bezüglich des Rückmeldekonzepts erfolgte eine Kritik der fehlenden Passung zu den hessischen Standards. Des Weiteren stand die mangelnde Individualisierung im Fokus der Gespräche. Dies erschwerte es den Befragten, gezielte Maßnahmen zur Weiterarbeit und Förderung abzuleiten. Zudem erhielten die Schüler aufgrund dessen eine sehr geringe bis keine Rückmeldung über ihre eigene Leistung. Trotz dieser Kritik äußerten einige Lehrkräfte und Schulleitungsvertreter den Wunsch, die Testergebnisse zu benoten, um die geleistete Arbeit auf diese Weise zu honorieren und der Lernstandserhebung einen größeren Stellenwert beizumessen.

Der Nutzen der Lernstandserhebung wurde primär im diagnostischen Potenzial für eine Stärken-Schwächen-Analyse gesehen. Der intendierte Mehrwert wurde erkannt, aber oftmals nicht in konkreten Handlungen und Maßnahmen umgesetzt. Des Weiteren wurde in der Lernstandserhebung ein Impuls zur Selbstreflexion gesehen, war wiederum eine internale Attribuierung beförderte. Auch der Nutzen für das Kennenlernen neuartiger Aufgabenformate wurde angeführt. Demgegenüber wurden die Funktionen des Testinstruments von einigen Befragten auch infrage gestellt, da die erhaltenen Informationen und Anregungen zuvor bereits bekannt gewesen seien. Auch sei der alltägliche Unterricht noch nicht auf die Kompetenzorientierung ausgelegt, so dass das Testkonzept noch nicht zum Unterrichtkon-

zept passe. Ein paar Befragte bewerteten weiterhin standardisierte Tests grundsätzlich als unsinnig für die Schulentwicklung. Insbesondere das Verhältnis zwischen Nutzen und dem dafür implizierten Zeitaufwand wurde von der Mehrheit der Befragten als negativ beurteilt, was bei fünf Personen beziehungsweise Schulen zu dem Entschluss führte, an keinem weiteren Testdurchlauf teilzunehmen.

Zuletzt wurden in der Auswertung die Probanden den Nutzungsformen zugeordnet. Am häufigsten lag bei 13 Befragten die symbolische Nutzung vor, welche sich primär durch die mangelhafte Umsetzung in der Aktionsphase auszeichnet. Demgegenüber konnte sechsmal eine instrumentelle Nutzung beobachtet werden, indem konkrete Handlungen und Maßnahmen sichtbar wurden. Davon wiesen drei Befragte einen besonderen Professionalisierungsgrad auf. In den anderen Fällen lag vorrangig eine konzeptionelle Nutzung vor. Ein Zusammenhang zwischen der Nutzungsform und dem getesteten Unterrichtsfach war nicht festzustellen. Bezogen auf eine Unterscheidung zwischen summativer und formativer Nutzung konnten nur fünf Befragte der formativen Nutzung zugeordnet werden, so dass die eindeutige Mehrheit der Gesprächspartner eine primäre summative Verwendung der Testergebnisse ohne Weiterentwicklung der Unterrichtsqualität verfolgt hat.

14 Beantwortung der Forschungsfragen

Nachdem die wesentlichen Untersuchungsergebnisse zusammenfassend dargestellt wurden, werden diese in einen gemeinsamen Kontext mit den theoretischen Betrachtungen geführt, um abschließend die Fragestellungen dieser Arbeit zu beantworten (vgl. Abschnitt 1.2).

Fragestellung 1: Welches Potenzial bieten die Vergleichsarbeiten für die Schulentwicklung?
Potenziale sind mögliche positive Effekte, die Entwicklungsprozesse im schulischen Kontext auslösen können. Dabei leiten sich die von den Landesinstituten und vom IQB intendierten Potenziale der Vergleichsarbeiten zunächst primär aus ihren zugrundeliegenden Funktionen ab.

Die Vergleichsarbeiten stellen als ein externes, standardisiertes Testinstrument neben anderen Diagnoseformen eine ergänzende Beurteilungsform der Schülerleistungen dar. Sie überprüfen, inwiefern die Schüler spezifische Fachkompetenzen erworben haben. Diese Zwischenkontrolle zu bestimmten Zeitpunkten der Schullaufbahn soll eine Bestandsaufnahme für verschiedene schulische Akteure - vorrangig für die unterrichtende Lehrkraft und den betreffenden Schüler sowie weiterführend für die Schulleitung und die Eltern - ermöglichen. Allerdings wird der Test unabhängig vom individuellen Unterricht konzipiert. Ein Rückschluss auf die Unterrichtsqualität ist problematisch, da eine Ursachenanalyse nicht in der wissenschaftlichen Testauswertung inbegriffen ist und durch die Lehrkraft selbst vorgenommen werden muss. Es stellt sich damit die Herausforderung, eine Verknüpfung zwischen den Testinhalten, den Inhalten des Unterrichts und den geförderten Kompetenzen herzustellen.

Zudem können in der Auswertung der Vergleichsarbeiten etwa bei geschlossenen Items die Lösungswege nicht berücksichtigt werden, da lediglich das Endergebnis analysiert wird. Die Reflexion dieser Items kann somit sowohl bei der standardisierten Auswertung als auch bei der Ursachenbetrachtung durch die Lehrkraft nur oberflächlich erfolgen. Eine umfassende Einschätzung des Kompetenzstandes ist aus diesem Grund nicht möglich. Ebenso ist eine absolute Durchführungs- und Korrekturobjektivität nicht gegeben, da die jeweilige Lehrkraft dies selbst vornimmt. Daher entsprechen die Testergebnisse nicht vollständig den Gütekriterien eines Tests, was die Aussagefähigkeit der Vergleichsarbeiten mindert.

Mithilfe von Referenzwerten, wie den Landesmittelwerten im Rahmen eines fairen Vergleichs, können die Testergebnisse einerseits sozial verortet und bewertet werden. Andererseits soll eine kriteriale Beurteilung auf Grundlage einer Zuordnung zu Kompetenzmodellen vorgenommen werden, so dass eine tiefgründige Analyse der Schülerleistungen ermög-

licht wird. Hierbei stellt sich das Problem, dass die Testaufgaben sowohl herkömmlichen Modellen wie den Anforderungsbereichen als auch den Kompetenzstufenmodellen zugewiesen werden. Eine kriteriale Bewertung der Testergebnisse ist damit erschwert, denn die verschiedenen Modelle sind zueinander nicht kompatibel.

Des Weiteren werden beispielsweise im hessischen Rückmeldungskonzept lediglich prozentuale Durchschnittsergebnisse angeführt. Eine Verortung zu den Kompetenzstufen unterbleibt hier vollständig. Es besteht die Frage, wie diese Durchschnittsprozentwerte im Hinblick auf die Erfassung eines Kompetenzstandes zu bewerten sind. Die kriteriale Auswertung im Sinne einer Analyse der Kompetenzausprägung ist daher nicht möglich, was die Aussagekraft der Vergleichsarbeiten enorm reduziert. Demgegenüber erfolgt in der Rückmeldung eine Schwerpunktsetzung auf den sozialen Vergleich mithilfe des korrigierten Landesmittelwerts. Dies verstärkt das Potenzial der Verortung, lenkt jedoch von der für die Unterrichtsentwicklung aufschlussreicheren kriterialen Analyse ab.

Darüber hinaus können die Vergleichsarbeiten mit einer Rechenschaftsablegung für die Lehrperson verknüpft sein, indem sich ihr Unterricht hinsichtlich seiner Wirksamkeit an den Testergebnissen teilweise - selbstverständlich nicht ausschließlich - messen lässt. Die Schülerresultate können weiterhin einen Teilaspekt der internen Schulevaluation darstellen, weil sie Aufschluss über die Qualität und Weiterentwicklung der Schülerleistungen geben.

Diese angeführten Potenziale sind rein summativer Natur und dem Bereich der Qualitätssicherung zuzuordnen. Demgegenüber bieten die Vergleichsarbeiten ebenso Potenziale formativer Art, die als Katalysator für die verschiedenen Bereiche der Schulentwicklung fungieren können.

Beispielsweise können auf Grundlage einer umfassenden Diagnose Reflexionen zu den Ursachen der Schülerleistungen sowie Stärken-Schwächen-Analysen vorgenommen werden. Im Sinne eines kompetenzorientierten Unterrichts können darauf aufbauend Förderziele formuliert und entsprechende Fördermaßnahmen initiiert werden. Die tatsächliche Umsetzung von konkreten Handlungen zur Qualitätsweiterentwicklung stellt dabei den Kern des gesamten Nutzungsprozesses der Vergleichsarbeiten dar. Für eine positive Wirkung der Vergleichsarbeiten ist es in diesem Zusammenhang von fundamentaler Bedeutung, dass die Testergebnisse ausschließlich zur Förderung und nicht zur Benotung verwendet werden. Eine individualisierte Diagnostik und Förderung ist durch zusätzliche Auswertungen der Lehrkraft zwar möglich, wird aber aufgrund eines zu großen Messfehlerrisikos in der standardisierten Testauswertung nicht berücksichtigt. Im hessischen Rückmeldekonzept werden dementsprechend nur Klassenwerte beziehungsweise Leistungsgruppen in Form von Quartilen ausgewiesen. Eine schülerspezifische Förderung im Sinne eines kompetenz-

orientierten Unterrichts ist in dessen Konsequenz enorm erschwert beziehungsweise nicht möglich.

Des Weiteren können die Tests aufgrund ihrer Inhalte, Aufgabenformate und Aufgabenkontexte Innovationsimpulse für die Unterrichtsgestaltung bieten. Bei einer zu starken Orientierung an geschlossenen Aufgabenformaten besteht die Gefahr der Etablierung eines Teaching to the Test. Dennoch tragen die Vergleichsarbeiten aufgrund ihrer Innovationsanreize sowie durch die Verwendung zusätzlicher didaktischer Materialien zu einer stärkeren Verankerung der Bildungsstandards sowie der Didaktik kompetenzorientierten Unterrichts im schulischen Alltag bei.

Die Vergleichsarbeiten sollen primär die Ausprägung vorhandener Kompetenzen im Kontext der Bildungsstandards überprüfen. Dies trifft jedoch nicht ohne Weiteres auf das untersuchte Bundesland Hessen zu. Dort wurden die Lernstandserhebungen zeitlich vor den Bildungsstandards implementiert, so dass eine Überprüfung der erreichten Kompetenzen entsprechend den Bildungsstandards bislang nicht erfolgen konnte. Selbstverständlich verfügen die Schüler über messbare Kompetenzen, aber eine Verknüpfung zwischen dem Unterricht, den Testergebnissen und den Bildungsstandards war nicht möglich. In diesem Fall ist das Innovationspotenzial umso bedeutsamer, indem die Lehrkräfte über die Vergleichsarbeiten verstärkt mit den Bildungsstandards konfrontiert werden.

Neben diesen Potenzialen für die Unterrichtsentwicklung kann auch die Organisationsentwicklung befördert werden. Die Kooperation und der kommunikative Diskurs über die Vergleichsarbeiten können sich innerhalb des Kollegiums intensivieren. Für den Bereich der Personalentwicklung ist die Stärkung der Lehrerprofessionalität zu nennen, da neue Impulse gewonnen und die diagnostischen Fertigkeiten ausgebaut werden.

Neben diesen Potenzialen für die Schulentwicklung üben die Vergleichsarbeiten weitere Funktionen im Bildungsmonitoring aus, wie die Bestandsaufnahme, die Feststellung schulischen Unterstützungsbedarfs, die Validierung der Kompetenzmodelle sowie die Erhöhung der Transparenz hinsichtlich der in den Bildungsstandards formulierten Anforderungen an die Schülerleistungen. Diese Aspekte sind in ihrer Bedeutung nicht minderwertiger. Da in der Untersuchung lediglich die Schulentwicklung betrachtet wurde, können zu diesen Funktionen jedoch sehr wenige Aussagen getroffen werden. Nach Information des LSA werden in Hessen diese Zielsetzungen mit den dortigen Lernstandserhebungen auch nicht verfolgt.

Fragestellung 2: Inwiefern wird das Potenzial der Vergleichsarbeiten in hessischen Gymnasien für die Schulentwicklung genutzt?

In der qualitativen Untersuchung in Form einer Interviewstudie mit Lehrkräften und Schulleitungsvertretern an zwölf hessischen Gymnasien konnte bei jedem Befragten eine Nutzung der Vergleichsarbeiten konstatiert werden. Jedoch verliefen die einzelnen Nutzungsphasen von stark unterschiedlicher Intensität. Während sich die Rezeption und die Reflexion der Testergebnisse in jedem Fall in verschiedenem Ausmaß vorgenommen wurden, waren konkrete Handlungen als Reaktionen auf die Schülerleistungen kaum bis gar nicht ersichtlich. Dies hatte Rückwirkungen auf die Ausprägung der sich anschließenden Evaluationsphase, die lediglich bei einem Befragten anhand einer Kontrollmaßnahme vorgenommen wurde. Die Phase der Aktion, in der konkrete Maßnahmen initiiert und umgesetzt werden, ist demnach die entscheidende Wendestelle im Nutzungsprozess, an der die Auseinandersetzung mit dem Test, der Rückmeldung und den zugehörigen Schülerleistungen oftmals rapide abnimmt beziehungsweise sogar abbricht.

Dabei ist deutlich zwischen vorhandener Handlungsbereitschaft aufgrund gewonnener Anregungen und der tatsächlichen Handlung zu differenzieren. Es konnte festgestellt werden, dass eine intensive Rezeption und Reflexion die Effektivität der Aktionsphase erhöht. Eine geringe Rezeption und Reflexion muss hingegen nicht automatisch einen Abbruch des Nutzungsprozesses zur Folge haben. Die Aktion stellt jedoch das Kernstück im Umgang mit den Vergleichsarbeiten dar, da in dieser Phase die Schulentwicklung konkret gefördert wird. Wenn die erfolgreiche Durchsetzung dieser Nutzungsphase nicht gelingt, ist es demnach mehr als fraglich, ob die Potenziale der Vergleichsarbeiten in der schulischen Realität überhaupt ihre Wirkung entfalten.

Um diese Fragestellung weiterführend zu beantworten, werden die erläuterten Potenziale erneut aufgegriffen und mit den Untersuchungsergebnissen in Beziehung gesetzt.

Die Möglichkeit einer Bestandsaufnahme der vorhandenen Fähig- und Fertigkeiten bei den Schülern wurde nur eingeschränkt genutzt. Alle Befragten nahmen die Ergebnisse ihrer Klasse zur Kenntnis und differenzierten diese teilweise nach fachlichen Kompetenzbereichen. Für eine tiefgründige Bestandaufnahme der Ausprägung des Kompetenzstandes ist jedoch eine Ursachenreflexion notwendig, welche sowohl internale als auch externale Attribuierungen umfasst. Für positive Ergebnisse wurden primär internale Ursachen angeführt, währenddessen bei defizitären Leistungen internale wie externale Attribuierungen vorgenommen wurden. Problematisch erwies sich in diesem Zusammenhang die starke Konzentration bei einem Großteil der Befragten auf den sozialen Vergleich anhand des hessischen korrigierten Landesmittelwerts. Wenn die Klasse in der Gesamtleistung und/ oder

in einzelnen Teilbereichen über dem Landesmittelwert abgeschnitten hatte, trat stets eine Zufriedenheit mit den Ergebnissen ein, die eine tiefgründige Ursachenanalyse verhinderte. Weiterführend wurde keine Notwendigkeit gesehen, konkrete Handlungen als Reaktion auf die Schülerleistung einzuleiten. Aus diesem Grund erfolgte die Bestandsaufnahme oftmals nur oberflächlich. Gefördert wurde dies des Weiteren durch eine mangelnde Auswertung nach Kompetenzstufen in der Rückmeldung, welche durch die Angabe durchschnittlicher Prozentwerte ersetzt wurde. Auch traten zum Teil Diskrepanzen zwischen den Testergebnissen und den vorherigen Unterrichtsbeobachtungen bei den Lehrkräften auf, was eine Bewertung der Schülerleistungen erschwerte.

Mit der Bestandsaufnahme geht das Potenzial der Verortung einher, welches von allen Befragten durchweg genutzt wurde. Gewährleistet wurde dies wiederum durch die starke Fokussierung auf den hessischen Referenzwert im Ergebnisbericht der Rückmeldung, was den sozialen Vergleich immens beförderte. Da es sich um Durchschnittswerte der Klasse, maximal differenziert nach Kompetenzbereichen, handelt, ist diese Verortung erneut nur als oberflächlich zu bewerten. Die Verortung erzeugte primär Zufriedenheit beziehungsweise Unzufriedenheit, was wie soeben beschrieben negative Konsequenzen für die Intensität des weiteren Nutzungsprozesses hatte.

Das Potenzial der Rechenschaftslegung wirkte sich erwartungsgemäß sehr selten aus, da die Lehrpersonen die Testergebnisse alleine oder im geschützten Rahmen des Fachkollegiums ausgewertet haben. Die Kommunikation mit der betreffenden Klasse beinhaltet keine Rechenschaftsfunktion für die Lehrperson, sondern eher für den individuellen Schüler. Mit den Eltern wurde nur in Einzelfällen über die Schülerleistung gesprochen, wobei in dem Gespräch der Fokus auf der weiteren Förderung lag und weniger auf dem Testresultat. Einzelgespräche zwischen der Lehrperson und der Schulleitung fanden in keinem Fall statt. Innerhalb der Schulleitung wurden die Ergebnisse allgemeiner und oberflächlicher besprochen, was nach Aussage einiger Lehrkräfte mit dem mangelnden Interesse der Schulleitungen zu begründen war. Demgegenüber benutzte die Schulleitung in zwei Fällen die Ergebnisse der Vergleichsarbeit zur Öffentlichkeitsarbeit, indem Zeitungsartikel über die Schulergebnisse erschienen. Dies zieht inoffizielle Vergleiche zwischen einzelnen Schulen automatisch nach sich, was den intendierten Funktionen der Leistungsmessung eindeutig konträr entgegensteht.

Innerhalb der Schulleitungsdiskussionen wurden teilweise die Leistungen einzelner Klassen miteinander verglichen, was jedoch kaum Rückschlüsse auf die Unterrichtsqualität der Schule erlaubte. Vergleiche über mehrere Testteilnahmen hinweg gab es nicht. Demnach sind die Vergleichsarbeiten kein Bestandteil des Schulevaluationskonzepts. Ein Grund hier-

für ist zum Teil die fehlende Passgenauigkeit zwischen den Testinhalten und den individuellen schulischen Unterrichtsschwerpunkten. Außerdem nehmen aufgrund der Freiwilligkeit in Hessen oftmals nicht alle Parallelklassen in einem Fach an dem Test teil, was Leistungsvergleiche auf Schulebene erschwert.

Diese genannten Potenziale sind auf der summativen Ebene der Schulentwicklung zu verorten und können bereits bei den Nutzungsphasen der Rezeption und Reflexion Wirkungen erzeugen. Das oftmals geringe Ausmaß der Aktion beziehungsweise das Fehlen einer Evaluation wirkte sich hingegen primär auf die formativen Potenziale der Vergleichsarbeiten aus, indem nur geringfügig Qualitätsentwicklungen initiiert und forciert wurden.

In diesem Zusammenhang ist das Potenzial des Diagnostizierens und Förderns zu nennen. Die Diagnose erfolgt bereits in der Rezeption und Reflexion und soll die Entwicklung eines Förderkonzepts zur Folge haben. Da allerdings wie erwähnt die Bestandsaufnahme und die Verortung der Schülerleistungen aufgrund der Fokussierung auf den sozialen Vergleich und der Angabe von Durchschnittswerten bei den Befragten eher oberflächlich verliefen, wurden so gut wie keine wirksamen Fördermaßnahmen ergriffen. Auch die in der Rückmeldung enthaltenen Anregungen zur Weiterarbeit leisteten aufgrund ihrer Kürze und standardisierten Formulierung keinen positiven Beitrag zur Umsetzung von Handlungen. Bei Zufriedenheit mit den Testergebnissen wurde von den Lehrkräften kein Handlungsbedarf erkannt. Zudem äußerten einige Befragte, dass sie durch die Rückmeldungen keine wesentlich neuen und aufschlussreichen Informationen erhalten hätten. Dies verhindert erneut die Aktionsphase.

Das Potenzial „Diagnostizieren und Fördern“ bezieht sich primär auf die Klassenebene. Nur wenige Lehrpersonen gaben ihren Schülern individualisiertes Feedback in Form von Kommentierungen im Aufgabenheft oder durch Gespräche. Die ergriffenen Maßnahmen von den Befragten stellten nie ein individualisiertes Fördern im Sinne des kompetenzorientierten Unterrichts dar. Die Lehrkräfte formulierten jedoch das Bedürfnis nach einer individualisierten Auswertung, was die Rückmeldung in keiner Weise fördere und daher nur mit einem erheblichen Mehraufwand zu leisten sei.

Des Weiteren sei erwähnt, dass einige Befragte die Vergleichsarbeit in selektiver Form benutzten, indem sie die Schülerleistungen nach Bewertungsmaßstäben benoteten, die der Testkonzeption nicht gerecht wurden. Im Kontext der unerwünschten Auswirkungen der Leistungsmessung zählen auch die Manipulationen, die von einigen Lehrpersonen vorgenommen wurden. Indem einzelne Schüler von der Auswertung ausgeschlossen wurden, Aufgabenteile nicht ausgewertet wurden und Unterschiede in der Vergabe der Punkte festzustellen waren, wurden die Ergebnisse der Schüler bewusst verfälscht, was die Objektivität

und somit die Aussagekräftigkeit der Testdaten enorm reduzierte. Da in der Untersuchung lediglich mit 19 Personen, die selbst die Vergleichsarbeit durchgeführt haben, Interviews stattfanden, ist das Ausmaß dieser Manipulationen unklar, aber wohl nicht zu unterschätzen.

Innovationsanreize durch die Testheftgestaltung wurden mehrfach in den Interviews festgestellt. Diese bezogen sich beispielsweise auf Ideen zu Aufgabenkontexten, der Schwerpunktlegung einzelner Kompetenzbereiche im Unterricht oder der Konzeption von internen Vergleichsklassenarbeiten. Die zur Verfügung gestellten Handreichungen, welche das Potenzial der Innovation in besonderem Maße befördern soll, wurde von keinem einzigen Befragten genutzt. Ihre Funktion wurde daher komplett verfehlt. Des Weiteren muss in diesem Zusammenhang erwähnt werden, dass die genannten Innovationsimpulse bei den Gesprächspartnern nicht in konkrete Handlungen mündeten. Zwar wurde die Schulentwicklung nicht direkt gefördert, jedoch kann das Wissen über diese neuartigen Möglichkeiten für die Unterrichtsgestaltung positive Konsequenzen für die Weiterentwicklung der Lehrerprofessionalität nach sich ziehen. Die weiteren in diesem Kontext intendierten Funktionen der verstärkten Orientierung an den Bildungsstandards sowie die Förderung der Medienkompetenz bei den Lehrpersonen wurden nicht erreicht.

Ebenso intensivierte sich die Kooperation in keiner untersuchten Schule. Die eher oberflächlichen Diskussionen beschränkten sich auf inoffizielle Gespräche und Fachkonferenzen. Neue Kooperationsstrukturen oder -gelegenheiten wurden dabei nicht geschaffen.

Wie bereits erläutert, werden in Hessen die Potenziale der Vergleichsarbeiten für das Bildungsmonitoring nicht gezielt verfolgt. Die konzipierten Handreichungen können jedoch auf die Verankerung der Bildungsstandards im Unterricht im Zuge ihrer Implementierung unterstützend einwirken. Da diese Manuale von keinem Befragten gelesen wurden, ist dieses Potenzial fehlgeschlagen. Zudem wurde von einem Befragten die fehlende Zuweisung zu den hessischen Zwischenstandards in den Klassenstufen 6 und 8 erwähnt, was der Funktion einer höheren Transparenz der in Bildungsstandards formulierten Anforderungen entgegensteht.

Um abschließend einen Überblick zu erhalten, inwiefern die Potenziale der Vergleichsarbeiten genutzt wurden, wird in dieser Stelle eine verkürzte Darstellung der intendierten Funktionen der Vergleichsarbeiten eingefügt und ihrer Wirkung gemäß eines Ampelprinzips farblich gekennzeichnet (vgl. Abbildung 21). Demnach bedeutet die Farbe Grün, dass die Nutzung dieses Potenzials bei allen Befragten festzustellen war. Bei Gelb wurde diese Funktion nur bei wenigen Befragten beziehungsweise in sehr abgeschwächter Form genutzt. Rot hingegen symbolisiert, dass die intendierte Funktion keine Wirkung zeigte.

Funktionen von Vergleichsarbeiten			
Individual-diagnostik	**Schul-und Unterrichtsentwicklung**		**Bildungsmonitoring**
	Qualitätssicherung (summativ)	Qualitätsentwicklung (formativ)	
Diagnostizieren und Fördern auf Schülerebene	Schulevaluation	Innovationsanreize	Bestandsaufnahme
Legende: Rot = Funktion zeigte keine Wirkung Gelb = Funktion zeigte nur eine abgeschwächte Wirkung Grün = Funktion wurde erfüllt	Bestandsaufnahme	Diagnostizieren und Fördern (auf Klassenebene)	Feststellung von schulischem Unterstützungsbedarf
	Verortung	Weiterentwicklung der Lehrerprofessionalität	Unterstützung der Implementierung von Bildungsstandards
	Rechenschaftslegung	Intensivierung der Fachgruppen- und Fachkonferenzarbeit	Transparenz der in den Bildungsstandards formulierten Anforderungen
			Validierung der Kompetenzmodelle

Abbildung 21: Nutzung der Potenziale der Vergleichsarbeiten an hessischen Gymnasien

Anhand der Abbildung 21 wird signifikant deutlich, dass lediglich das Potenzial der Verortung durch die Vergleichsarbeiten umfassend genutzt wurde. Dies regt jedoch keine aktive Schulentwicklung an, da sie dem Bereich der Qualitätssicherung und nicht der Qualitätsentwicklung zu verorten ist. Dass die Potenziale für eine formative Schulentwicklung nicht beziehungsweise lediglich in abgeschwächter Form zutage treten, ist für den tatsächlichen Mehrwert der Vergleichsarbeiten für die schulische Arbeit äußerst problematisch. Da die Lehrkräfte vor der Testdurchführung spezifische Erwartungen an den Nutzen dieser Leistungsmessung stellen, sind aufgrund der geringen Verwirklichung ihrer intendierten Funktionen Enttäuschungen nicht unwahrscheinlich. Solche Diskrepanzen zwischen den Erwartungen und den späteren Effekten konnten auch in der Untersuchung beobachtet werden.

Fragestellung 3: Welche Faktoren greifen fördernd beziehungsweise hemmend in den Nutzungsprozess durch die Lehrkräfte ein?

Wie aus dem Zyklenmodell von Helmke (Helmke A. , 2004) hervorgeht, wird der Nutzungsprozess der Vergleichsarbeiten durch verschiedene Bedingungen beeinflusst, welche sich in die drei Oberkategorien individuelle, schulische und externe Bedingungen kategorisieren lassen. Es ist ein weiteres Ziel der Fragestellung in dieser Untersuchung, die zu beobachtenden Einflussfaktoren mit den von Helmke genannten zu vergleichen und das Ausmaß ihres Einflusses auf die Nutzung zu bewerten. Da das Zyklenmodell von Helmke als Grundlage für die Auswertung der Interviews diente, wurden die Einflussfaktoren jeder einzelnen

Nutzungsphase zunächst untersucht und in den Abschnitten 9 und 10 beschrieben. An dieser Stelle werden aus diesen einzelnen Aspekten phasenübergreifende Einflussfaktoren herausgefiltert welche auf den gesamten Nutzungsprozess einwirken. Ihr Einfluss ist dementsprechend immens und muss bei der Bewertung der tatsächlich genutzten Potenziale der Vergleichsarbeiten (vgl. Fragestellung 2) berücksichtigt werden.
Die Abbildung 22 stellt die gravierendsten Einflussfaktoren graphisch dar.

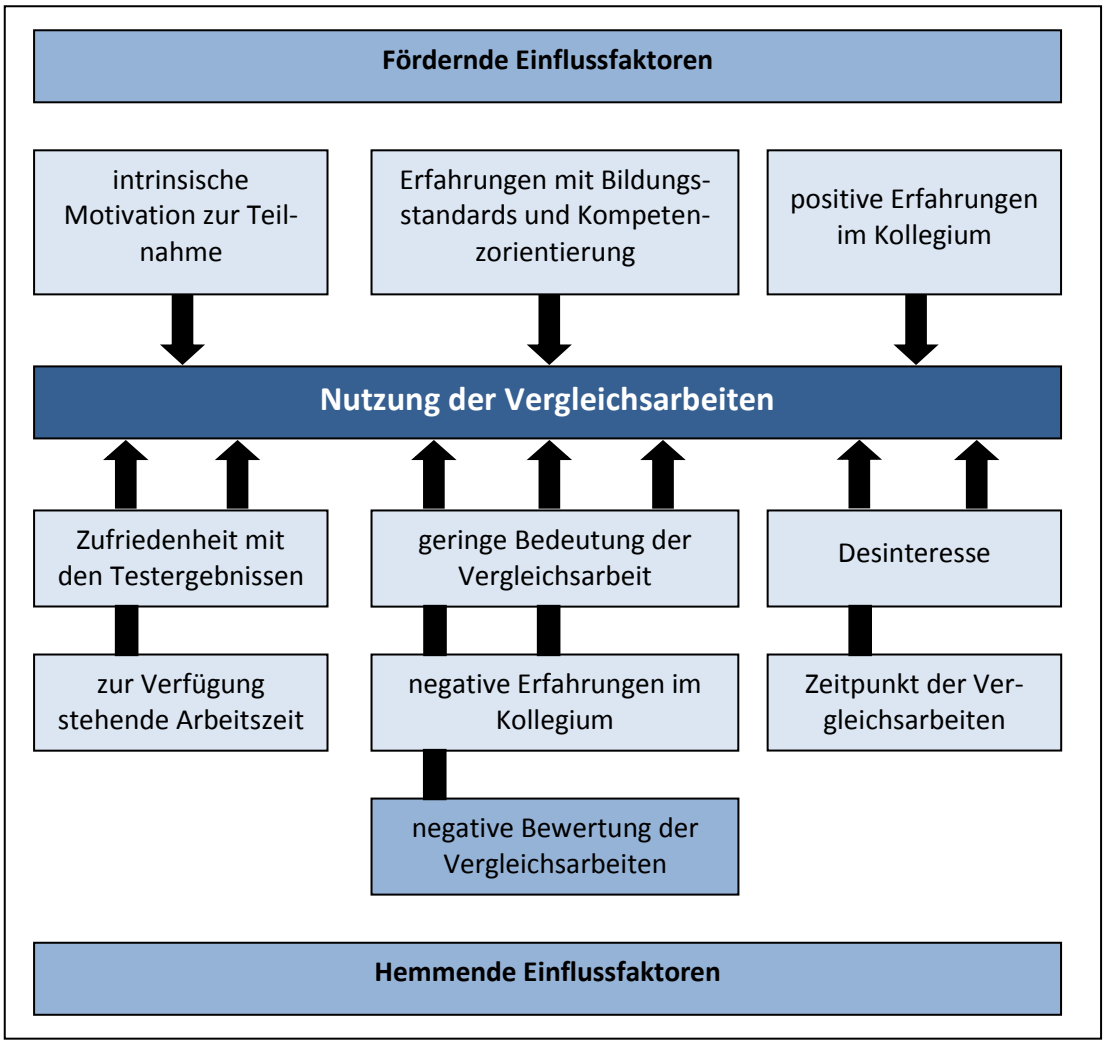

Abbildung 22: Einflussfaktoren auf den Nutzungsprozess

Ein positiver individueller Einflussfaktor ist die *intrinsische Motivation* zur Teilnahme an den Vergleichsarbeiten. Dies ist primär dem Konzept der Freiwilligkeit in Hessen zu verdanken. Jedoch trifft diese intrinsische Motivation nicht auf jede Lehrkraft zu, da die Schulleitungen die Teilnahme in ihrer Schule verbindlich festlegen können. Die selbstständige Entscheidung zur Durchführung des Tests setzt sowohl ein gewisses Interesse als auch eine Erwartungshaltung an den Sinn der Vergleichsarbeiten voraus. Wie sich in der Interviewstudie gezeigt hat, wirkt sich dies auf die Intensität der Arbeitsbereitschaft im Umgang mit den Tests und

den Rückmeldungen positiv aus. Die Nutzung erfolgt tiefgründiger und hat verstärkt die Umsetzung konkreter Maßnahmen zur Folge.

Als ein weiterer bedeutender Aspekt ist der *Professionalisierungsgrad* im Kontext der Bildungsstandards und der Didaktik des kompetenzorientierten Unterrichts zu nennen. Einige Lehrpersonen hatten bereits mehr Erfahrungen in diesem Bereich durch ihre Arbeit an Universitäten oder in Studienseminaren bei der Lehrerausbildung und richteten ihren Unterricht verstärkt daraufhin aus. Folglich konnte das Potenzial der Testaufgaben im Kontext zu den Bildungsstandards mit ihren zugehörigen Kompetenzen stärker Wirkung zeigen. Diese Lehrpersonen wiesen durchgängig eine intensivere Nutzung der Vergleichsarbeiten in den Phasen der Rezeption bis Aktion auf. Dies festigt die Vermutung, dass die Potenziale der Vergleichsarbeiten stärker zum Tragen kommen werden, wenn die Bildungsstandards in der Unterrichtsgestaltung mehr verankert sind.

Ein weiterer Einflussfaktor auf Schulebene sind die bisherigen *Erfahrungen* innerhalb des Kollegiums. Sind sie positiv, wirkt sich dies zum einen auf die intrinsische Motivation der Lehrkräfte aus, was wiederum eine verstärkte Teilnahme an den Vergleichsarbeiten zur Folge hat. Zum anderen verstärkt es die Kommunikation innerhalb der Lehrerschaft über die Tests und den zugehörigen Erfahrungen. Andererseits kann dieser Aspekt enorme negative Folgen nach sich ziehen, wenn die bislang gewonnenen Erfahrungen als ergebnislos bewertet wurden. Es entstehen bereits vor der Durchführung der Tests Ressentiments gegenüber den Vergleichsarbeiten, was sich massiv auf die Arbeitsbereitschaft und die Erwartungshaltung auswirkt.

Dies kann in Verbindung zu einem weiteren negativen Einflussfaktor stehen, dem *Desinteresse*. Wenn bereits zu Beginn die individuelle Motivation zur Durchführung und Korrektur der Vergleichsarbeiten sehr gering ist, verläuft die Nutzung der Testergebnisse in einem stark eingegrenzten und oberflächlichen Rahmen. Befördert wird dies durch den *geringen Stellenwert*, der dieser Leistungsmessung in einigen Schulen eingeräumt wird. Indem die Testergebnisse und deren Analyse ohne Folge bleiben, wird ihnen im Schulalltag keine besondere Bedeutung beigemessen. Daher äußerte eine Vielzahl der Lehrkräfte in den Interviews den Wunsch, die Testergebnisse benoten zu können, um den Stellenwert auf diese Weise zu erhöhen. Des Weiteren wird dies verstärkt durch die Wahrnehmung der Lehrpersonen, für eine intensive Nutzung der Vergleichsarbeiten nicht genügend *Arbeitszeit* zur Verfügung zu haben. Dem wirkten einige Schulen entgegen, indem Korrekturtage bewilligt wurden, was positive Folgen für die Einschätzung des Arbeitsaufwandes und für die Bereitschaft zur Analyse der Testergebnisse hatte.

In dieses komplexe Wirkungsgeflecht der Einflussfaktoren gehört ebenso der *Testzeitpunkt*, welcher besonders im Gymnasium mit der Durchführung des Abiturs kollidiert. Gerade in dieser Phase des Schuljahres ist der empfundene Stresspegel bei den Lehrkräften besonders hoch, so dass die Bereitschaft sich mit zusätzlichen und freiwilligen Tests zu beschäftigen, welche nach eigener Einschätzung nur eine geringe Bedeutung für die eigene Arbeit haben, gering ist. Zudem werden die ausführlichen Testergebnisse mit der Rückmeldung erst Ende Mai zur Verfügung gestellt. Das Schuljahr ist innerhalb weniger Wochen beendet, was den Handlungsspielraum für eine Nutzung der Schülerresultate im Sinne einer langfristigen Leistungsförderung der Klasse enorm reduziert.

Die Zufriedenheit mit den Testergebnissen beeinflusst ebenfalls die Nutzungsintensität und führt dazu, dass einige Funktionen der Vergleichsarbeiten nicht umgesetzt werden. Allerdings muss angemerkt werden, dass diese Zufriedenheit zumindest die Durchführung einer Rezeption der Resultate voraussetzt. Indem die Ergebnisse als positiv bewertet werden, wird kein Handlungsbedarf empfunden. Folglich wird die Reflexion oberflächlich vorgenommen und die Initiierung von Maßnahmen unterbleibt oft vollständig. Erzeugt wird die Zufriedenheit primär durch die Fokussierung auf den sozialen Vergleich in der Rückmeldung, bei dem die Klassenresultate mit dem korrigierten Referenzmittelwert verglichen werden.

Abgesehen von der Bedeutung der intrinsischen Motivation und dem Professionalisierungsgrad sind die genannten Einflussfaktoren im Zyklenmodell von Helmke (Helmke A. , 2004) nicht vorhanden. Der bedeutendste Aspekt stellt jedoch die Bewertung der Vergleichsarbeiten dar, welcher massiv auf den gesamten Nutzungsprozess einwirkt. Diese individuelle Bewertung baut sich bereits mit der Testdurchführung auf, indem die Einschätzungen zu der Testgestaltung, den Aufgabeninhalten, der Aufgabenschwierigkeit sowie dem Umgang der Schüler mit dem Testheft reflektiert werden. Die Erfahrungen bei der Korrektur und den zugrundeliegenden Vorgaben konkretisieren dieses Bild. Der Inhalt, die Aufbereitung und Aussagekraft der Rückmeldung werden anschließend ebenfalls in der Bewertung berücksichtigt. Schließlich kann zum Abschluss des Nutzungsprozesses von der Lehrkraft ein umfassendes Urteil über die Qualität und dem persönlichen Mehrwert der Vergleichsarbeit formuliert werden. Dies wirkt sich wiederum auf die Motivation für eine weitere Testdurchführung aus.

Sobald ein oder mehrere Aspekte dieser Bewertungskette negativ beurteilt werden, reduziert sich die Bereitschaft, sich mit der Vergleichsarbeit auseinander zu setzen, was eine Abschwächung der Nutzung und weiterführend eine Verringerung der Potenzialentfaltung zur Folge hat. Letztlich kann von den schulischen Akteuren nicht erwartet werden, dass sie

sich intensiv mit der Leistungsmessung auseinandersetzen, wenn sie deren Qualität als nicht ausreichend bewerten. Von einer Vielzahl der Befragten wurden in diesem Kontext massive Kritik am Test- und Rückmeldekonzept geäußert, was eine zentrale Ursache für die starke Differenz zwischen den intendierten Funktionen und dem tatsächlich eingetretenen Nutzen darstellt. Aus diesem Grund sollte die Bewertung der Vergleichsarbeiten durch die Lehrkraft als eine wesentliche Bedingung für den Nutzungsprozess im Zyklenmodell von Helmke (Helmke A. , 2004) ergänzt werden (vgl. Abbildung 23).

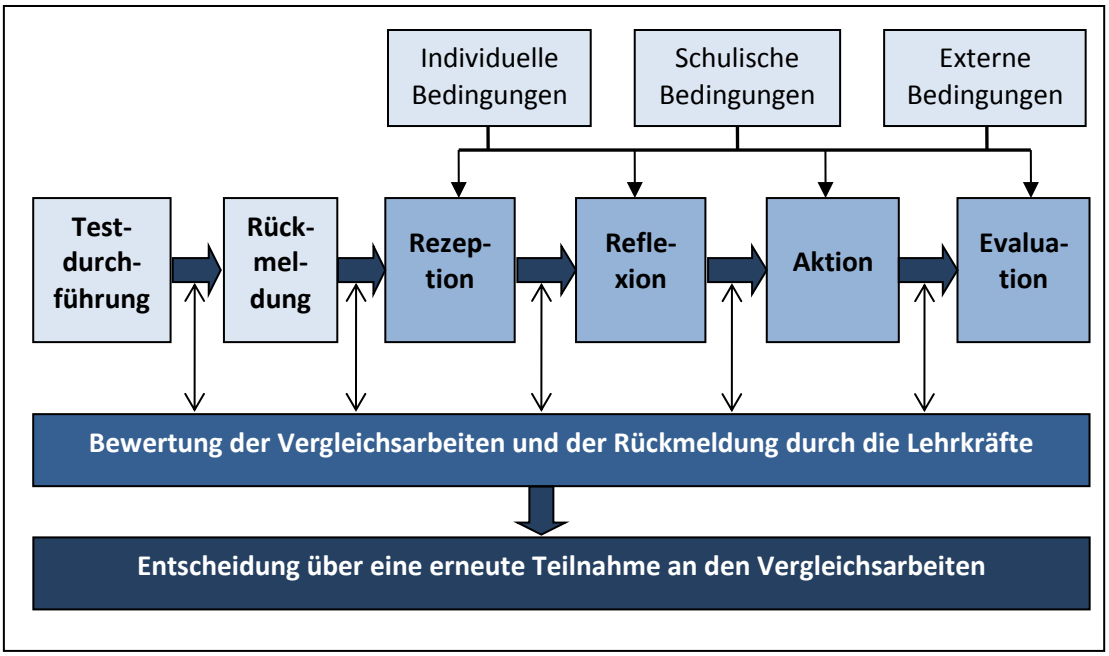

Abbildung 23: Erweiterung des Zyklenmodells von Helmke um den Einflussfaktor der Bewertung

Aus der Beantwortung der drei Fragestellungen dieser Arbeit werden zwei zentrale Aspekte deutlich: Zum einen ermöglichen die Vergleichsarbeiten bislang keine eindeutige und umfassende Kompetenzdiagnostik. Indem lediglich Durchschnittsprozentwerte angegeben werden, ist eine kompetenzorientierte Auswertung nicht möglich. Die kriteriale Interpretation der Schülerleistungen kann nicht erfolgen, da entsprechende Bewertungsmaßstäbe fehlen zum Beispiel in Form einer Zuordnung zu Kompetenzstufen fehlen. Die Lehrkräfte bräuchten jedoch eine konkrete Kompetenzdiagnostik für eine formative Nutzung der Vergleichsarbeiten und dies im Idealfall für individuelle Schüler. Die Konzentration der Rückmeldung auf den sozialen Vergleich befördert lediglich die summative Verwendung der Tests. Dass die Ableitung von spezifischen Fördermaßnahmen für einzelne Lernende beziehungsweise Schülergruppen ohne eine enorme Eigenleistung der Lehrkraft nicht möglich ist, steht dem Konzept des kompetenzorientierten Unterrichts vehement entgegen.

Dies korreliert mit dem zweiten wesentlichen Fazit. Die vorhandenen Potenziale der Vergleichsarbeiten werden nicht ausreichend genutzt. Dies wird einerseits dadurch gefördert, dass Gütekriterien (zum Beispiel die Auswertungsobjektivität) sowie die Anforderungen an eine Rückmeldung (zum Beispiel Relevanz, Bezugsnormen und Zeitnähe) nicht gewährleistet werden. Andererseits werden den Vergleichsarbeiten zu viele Funktionen zugewiesen, die unmöglich gänzlich verwirklicht werden können. Die Funktionen ergänzen sich zum Teil; einige widersprechen sich auch und führen zu falschen Erwartungen bei den Lehrkräften, die anschließend bei der Nutzung enttäuscht werden. Dies wirkt sich negativ auf die Nutzungsintensität bei den Lehrpersonen sowie auf die zukünftige Bereitschaft zur Teilnahme aus und führt dazu, dass die Vergleichsarbeiten ihren Funktionen in der Praxis nicht wie gewünscht gerecht werden.

Diese Leistungsmessung ist letztlich ein Diagnoseinstrument mit Grenzen in ihrer Aussagekraft, welche neben anderen Leistungsmessungsmöglichkeiten einsetzbar ist. Allerdings ist aufgrund der Ergebnisse dieser Untersuchung anzuzweifeln, ob derzeit ihr tatsächlicher Nutzen für die Schulentwicklung den enormen Kosten- und Arbeitsaufwand, der mit den Vergleichsarbeiten auf allen Ebenen des Bildungssystems verbunden ist, rechtfertigt.

15 Ausblick

Die in den Vergleichsarbeiten implizierten Potenziale wurden zum Untersuchungszeitpunkt laut der Auswertung der durchgeführten Interviews noch nicht ausreichend genutzt. Wie aufgezeigt wurde, ist dies zum Teil mit der Implementierung der Vergleichsarbeiten vor der Einführung der Bildungsstandards in Hessen zu begründen. Aus diesem Grund kann vermutet werden, dass bei zunehmender Etablierung der Bildungsstandards im Schulalltag sowie bei der Gestaltung des Unterrichts nach kompetenzorientierten Merkmalen die Vergleichsarbeiten ein größeren Nutzen für die Lehrkräfte entwickeln werden. Inwiefern die Ergebnisse dieser Untersuchung sich bei weiteren Testdurchläufen verändern, sollte demnach wissenschaftlich betrachtet werden.

Des Weiteren ist bei den Resultaten dieser Arbeit zu berücksichtigen, dass die Lehrkräfte in den Interviews oftmals von Anregungen für die Weiterarbeit sprachen. Es ist möglich, dass diese Ideen erst zu einem späteren Zeitpunkt umgesetzt werden, wenn beispielsweise dieselbe Jahrgangsstufe erneut unterrichtet wird. In diesem Fall müsste der Effekt der Vergleichsarbeiten rückblickend wesentlich positiver beurteilt werden. Aufgrund des gewählten Untersuchungsdesigns konnte dieser Aspekt in dieser Arbeit nicht erfasst werden. Eine Langzeitstudie wäre daher notwendig, welche die Auswirkungen der Erfahrungen und der abschließenden Beurteilung der Vergleichsarbeiten durch die Lehrkraft auf ihre nächste Testteilnahme untersucht.

Da die vielfältigen Funktionen der Vergleichsarbeiten eine Erwartungshaltung bei den schulischen Akteuren befördert, welche zum Teil nicht erfüllt wird, sollte dieser Aspekt und dessen Einfluss auf den Nutzen der Leistungsmessung weiterführend analysiert werden.

Ebenso konnte die Bedeutung des Einflussfaktors „Bewertung der Vergleichsarbeiten“ auf die Nutzungsintensität herausgestellt werden. Weil die vorliegende Untersuchung lediglich an zwölf hessischen Gymnasien vorgenommen wurde, ist dies in Form einer quantitativen Studie zu spezifizieren.

Literaturverzeichnis

Ackermann, H., & Rosenbusch, H. S. (2002). Qualitative Forschung in der Schulpädagogik. In E. König, & P. Zedler, *Qualitative Forschung. Grundlagen und Methoden* (2. überarb. Aufl. Ausg., S. S. 31-54). Weinheim, Basel: Beltz Verlag.

Altrichter, H. (2010). Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In H. Altrichter, & K. Maag Merki, *Handbuch Neue Steuerung im Schulsystem* (S. 219-254). Wiesbaden: VS Verlag für Sozialwissenschaften.

Altrichter, H., & Rürup, M. (2010). Schulautonomie und die Folgen. In H. Altrichter, & K. Maag Merki, *Handbuch Neue Steuerung im Schulsystem* (S. 111-144). Wiesbaden: VS Verlag für Sozialwissenschaften.

Amrein, A. L., & Berliner, D. C. (2002). High-Stakes-Testing, Uncertainty, and Student Learning. *Education Policy Analysis Archives , 10* (18).

Arnold, K.-H. (2001). Qualitätskriterien für die standardisierte Messung von Schulleistungen. Kann eine (vergleichende) Messung von Schulleistungen objektiv, repräsentativ und fair sein? In F. E. Weinert, *Leistungsmessungen in Schulen* (S. 117-130). Weinheim: Beltz.

Artelt, C. (2007). Externe Evaluation und einzelschulische Entwicklung - Ein zukunftsreiches Entwicklungsverhältnis? In J. Van Buer, & C. Wagner, *Qualität von Schule. Ein kritisches Handbuch* (S. 131-140). Frankfurt am Main: Lang.

Artelt, C., & Riecke-Baulecke, T. (2004). *Bildungsstandards: Fakten, Hintergründe, Praxistipps.* München: Oldenbourg.

Avenarius, H., Ditton, H., Döbert, H., Klemm, K., Klieme, E., Rürup, M., et al. (2003). *Bildungsbericht für Deutschland. Erste Befunde.* Opladen: Leske und Budrich.

Bähr, K. (2006). Erwartungen von Bildungsadministrationen an Schulleistungstests. In H. Kuper, & J. Schneewind, *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 127-141). Münster, New York, München, et al.: Waxmann.

Ballasch, H. (2009). Aus Vergleichsarbeiten lernen. In T. Bohl, & H. Kiper, *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 299-304). Bad Heilbrunn: Klinkhardt.

Bartnitzky, H. (2006). Wie VERA und Verwandtes die Bildungsqualität beschädigen. Die potemkinschen Dörfer der gegenwärtigen Schulpolitik. *Die Deutsche Schule , 98* (2), S. 201-213.

Bauer, K.-O. (2009a). Professionelles Selbst und Evaluation. In K.-O. Bauer, & N. Logemann, *Kompetenzmodelle und Unterrichtsentwicklung* (S. 75-112). Bad Heilbrunn: Klinkhardt.

Bauer, K.-O. (2009b). Professionelles Selbst, Evaluieren und Innovieren. In T. Bohl, & H. Kiper, *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 219-238). Bad Heilbrunn: Klinkhardt.

Baumert, J., Bos, W., & Lehmann, R. (2000). *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie - Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn, Bd. 1, Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit.* Opladen: Leske + Budrich.

Baumert, J., Bos, W., & Watermann, R. (1999). *TIMSS/III. Schülerleistungen in Mathematik und den Naturwissenschaften am Ende der Sekundarstufe II im internationalen Vergleich. Zusammenfassung deskriptiver Ergebnisse* (2. Ausg.). Berlin: Max-Planck-Institut für Bildungsforschung.

Boes, A. (2003). Tests - Sinn und Unsinn. Eine kritische Betrachtung. *PÄD-Forum: unterrichten erziehen , 31* (2), S. 94-97.

Bohl, T., Kleinknecht, M., & Maier, U. (2008). Datenbasierte Selbst- und Fremdevaluation. Eine exemplarische Analyse des Steuerungskonzeptes in Baden-Württemberg. *Die Deutsche Schule , 100* (4), S. 459-466.

Böhme, K. (2006). Testen: ja - Den Unterricht verarmen: nein. Aus Erfahrungen anderer Länder lernen. *Grundschule , 38* (5), S. 8-10.

Bonsen, M. (2010). Schulleitungshandeln. In H. Altrichter, & K. Maag Merki, *Handbuch Neue Steuerung im Schulsystem* (S. 277-294). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bonsen, M., & von der Gathen, J. (2004). Schulentwicklung und Testdaten. Die innerschulische Verarbeitung von Leistungsrückmeldungen. *Jahrbuch der Schulentwicklung , 13*, S. 225-252.

Bonsen, M., Büchter, A., & Peek, R. (2006). Datengestützte Schule- und Unterrichtsentwicklung. Bewertungen der Lernstandserhebungen in NRW durch Lehrerinnen und Lehrer. *Jahrbuch der Schulentwicklung , 14*, S. 125-148.

Bonsen, M., von der Gathen, J., Iglhaut, C., & Pfeiffer, H. (2002). *Die Wirksamkeit von Schulleitung. Empirische Annäherungen an ein Gesamtmodell schulischen Leitungshandelns.* Weinheim: Juventa.

Bortz, J., & Döring, N. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3. Ausg.). Berlin: Springer.

Bos, W., & Voss, A. (2008). Empirische Schulentwicklung auf Grundlage von Lernstandserhebung. Ein Plädoyer für einen reflektierten Umgang mit Ergebnissen aus Leistungstests. *Die Deutsche Schule , 100* (4), S. 449-458.

Böttcher, W. (2003). Bildung, Standards, Kerncurricula. Ein Versuch, einige Missverständnisse auszuräumen. *Die Deutsche Schule , 95* (2), S. 152-164.

Böttcher, W. (2009). Outputsteuerung durch Bildungsstandards. In H. Buchen, & H.-G. Rolff, *Professionswissen Schulleitung* (2. Ausg., S. 673-710). Weinheim, Basel: Beltz.

Böttcher, W. (2008). Standards. Konsequenzen der Output-Steuerung für die Lehrerprofessionalität. In W. Helsper, S. Busse, M. Hummrich, & R.-T. Kramer, *Pädagogische Professionalität in Organisationen. Neue Verhältnisbestimmungen am Beispiel der Schule* (S. 187-203). Wiesbaden: VS Verlag für Sozialwissenschaften.

Breiter, A., & Stauke, E. (2007). Anforderungen an elektronische Rückmeldesysteme aus Nutzersicht. *Empirische Pädagogik , 21* (4), S. 383-400.

Brinkmann-Hein, D., & Reh, S. (2005). Der Arbeitsplatz von LehrerInnen: Welche Rolle spielen Kooperation und professionelle Reflexion? *Journal für Schulentwicklung , 9* (2), S. 30-36.

Buchen, H. (2009). Schule managen - statt nur verwalten. In H. Buchen, & H.-G. Rolff, *Professionswissen Schulleitung* (2. Ausg., S. 12-101). Weinheim, Basel: Beltz.

Büchter, A., & Leuders, T. (2005). Zentrale Tests und Unterrichtsentwicklung... bei guten Aufgaben und gehaltvollen Rückmeldungen kein Widerspruch. *Pädagogik , 57* (5), S. 14-18.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2 Ausg.). München: Pearson Studium.

Buhren, C. G., & Rolff, H.-G. (2000). Personalentwicklung als Beitrag zur Schulentwicklung. *Jahrbuch der Schulentwicklung , 11*, S. 257-296.

Burkard, C. (1995). *Selbstevaluation. Ein Beitrag zur Qualitätsentwicklung von Einzelschulen?* Bönen: Verlag für Schule und Weiterbildung, Kettler.

Criblez, L., Oelkers, J., Reusser, K., Berner, E., Halbheer, U., & Huber, C. (2009). *Bildungsstandards.* Seelze-Velber: Kallmeyer Klett.

Crooks, T. J. (1988). The Impact of Classroom Evaluation Practices on Students. *Review of Educational Research , 58* (4), S. 438-481.

Dedering, K., Kneuper, D., Kuhlmann, C., Nessel, I., & Tillmann, K.-J. (2007). Bildungspolitische Aktivitäten im Zuge von PISA - das Beispiel Bremen. Zur politischen Legitimationskraft einer Leistungsvergleichsstudie. *Die Deutsche Schule , 99* (4), S. 408-421.

Demmer, M., & Schweitzer, J. (2005). Es fährt ein Zug nach nirgendwo... Zwischenbilanz einer unaufhaltsamen (?) Entwicklung. *Friedrich Jahresheft , 23*, S. 68-69.

Deutscher Bildungsrat. (1973). *Empfehlungen der Bildungskommission. Zur Reform von Organisation und Verwaltung im Bildungswesen. Teil 1: Verstärkte Selbstständigkeit der Schule und Partizipation der Lehrer, Schüler und Eltern.* Bonn: Klett.

Ditton, H. (2000a). Elemente eines Systems der Qualitätssicherung im schulischen Bereich. In H. Weishaupt, *Qualitätssicherung im Bildungswesen. Problemlage und aktuelle Forschungsbefunde* (S. 13-25). Erfurt: Pädagogische Hochschule Erfurt, Institut für Allgemeine Erziehungswissenschaft und Empirische Bildungsforschung.

Ditton, H. (2000b). Qualitätskontrolle und -sicherung in Schule und Unterricht. Ein Überblick über den Stand der empirischen Forschung. In A. Helmke, W. Hornstein, & E. Terhart, *Qualität und Qualitätssicherung im Bildungsbereich: Schule, Sozialpädagogik, Hochschule* (S. 73-92). Weinheim, Basel: Beltz.

Dobbelstein, P., & Peek, R. (2004). *Lernstandserhebungen in Nordrhein-Westfalen. Entwicklungsstand und Perspektiven. Präsentation auf der 1. EMSE-Fachtagung im Dezember 2004 in Soest.* Abgerufen am 11. November 2010 von http://www.emse-netzwerk.de/uploads/Main/EMSE01.zip

Drieschner, E. (2009). *Bildungsstandards praktisch. Perspektiven kompetenzorientierten Lehrens und Lernens.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Ehmke, T., Leiß, D., Blum, W., & Prenzel, M. (2006). Entwicklung von Testverfahren für die Bildungsstandards Mathematik. Rahmenkonzeption, Aufgabenentwicklung, Feld- und Haupttest. *Unterrichtswissenschaft , 34* (3), S. 220-238.

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, burteilen.* Abgerufen am 9. Januar 2011 von http://commonweb.unifr.ch/pluriling/pub/cerleweb/portfolio/downloadable-docu/Referenzrahmen2001.pdf

Fend, H. (1986). "Gute Schulen - schlechte Schulen". Die einzelne Schule als pädagogische Handlungseinheit. *Die Deutsche Schule , 78* (3), S. 275-293.

Fengler, J. (2009). *Feedback geben. Strategien und Übungen* (4. Ausg.). Weinheim, Basel: Beltz.

Fleischer, J., Spoden, C., Wirth, J., & Leutner, D. (2008). Flächendeckende Lernstandserhebungen - spezifische Herausforderungen und Lösungsansätze. Das Beispiel lernstand 8 in Nordrhein-Westfalen. In W. Böttcher, W. Bos, H. Döbert, & H. G. Holtappels, *Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive. Dokumentation zur Herbsttage der Kommission*

Bildungsorganisation, -planung, -recht (KBBB) (S. 195-207). Münster, New York, München, et al.: Waxmann.

Frey, A. (2008). Adaptives Testen. In H. Moosbrugger, & A. Kelava, *Testtheorie und Fragebogenkonstruktion* (S. 261-278). Heidelberg: Springer.

Frey, A., & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. In M. Prenzel, I. Gogolin, & H.-H. Krüger, *Kompetenzdiagnostik* (S. 169-184). Wiesbaden: VS Verlag für Sozialwissenschaften.

Fussangel, K., Rürup, M., & Gräsel, C. (2010). Lehrerfortbildung als Unterstützungssystem. In H. Altrichter, & K. Maag Merki, *Handbuch Neue Steuerung im Schulsystem* (S. 327-354). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gasteiger, H. (2007). Orientierungsarbeiten. Chancen nutzen mit Blick auf das Kind und den eigenen Unterricht. *Die Grundschulzeitschrift , 21* (207), S. 28-33.

Gaupp, N. (2008). Computerbasierte Erfassung sozialer Kompetenzen mit Video-Vignetten. In N. Jude, J. Hartig, & E. Klieme, *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte und Methoden* (S. 73-80). Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Gehrmann, A. (2003). *Der professionelle Lehrer. Muster der Begründung. Empirische Rekonstruktion.* Opladen: Leske und Budrich.

Gewerkschaft Erziehung und Wissenschaft - Landesverband Hessen. (15. Juni 2010). *Bildungsstandards - Kerncurricula für Hessen - GEW-Landesvorstand moniert "gravierende Mängel".* Abgerufen am 10. Mai 2011 von http://www.gew-hessen.de/index.php?id=296&tx_ttnews[backPid]=632&tx_ttnews[pS]=1286433198&tx_ttnews[pointer]=13&tx_ttnews[tt_news]=4488&cHash=4c51645e4c6e6b06eb1e6d8b68cb5b88

Gläser, J., & Laudel, G. (2010). *Experteninterviews und qualitative Inhaltsanalyse* (4. Aufl. Ausg.). Wiesbaden: VS Verlag.

Granzer, D. (2008). Bildungsqualität entwickeln durch Implementation und Evaluation von Standards. In D. Granzer, & P. Wendt, *Selbstevaluation in Schulen. Theorie, Praxis und Instrumente* (S. 49-61). Weinheim, Basel: Beltz.

Granzer, D. (2006). Von guten und "anderen" Aufgaben. *Grundschule , 38* (5), S. 18-20.

Gräsel, C., Fußangel, K., & Pröbstel, C. (2006). Lehrkräfte zur Kooperation anregen - eine Aufgabe für Sisyphos? *Zeitschrift für Pädagogik , 52* (2), S. 205-219.

Groß Ophoff, J., Hosenfeld, I., & Koch, U. (2007). Formen der Ergebnisrezeption und damit verbundene Schul- und Unterrichtsentwicklung. *Empirische Pädagogik , 21* (4), S. 411-427.

Groß Ophoff, J., Koch, U., Hosenfeld, I., & Helmke, A. (2006). Ergebnisrückmeldungen und ihre Rezeption im Projekt VERA. In H. Kuper, & J. Schneewind, *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 19-40). Münster, New York, München, et al.: Waxmann.

Grunder, H.-U. (2004). Schulentwicklung und die Konsequenzen für Schule und Unterricht. In R. Arnold, & C. Griese, *Schulleitung und Schulentwicklung. Voraussetzungen, Bedingungen, Erfahrungen* (S. 73-90). Hohengehren: Schneider-Verlag.

Haenisch, H., & Müller, S. (2005). Wann gelingen Parallelarbeiten und was bewirken sie? Eine qualitative Studie. *Die Deutsche Schule , 97* (3), S. 302-314.

Hallinger, P., & Heck, R. H. (1995). The principal's role in school effectiveness: An assessment of methodological progress, 1980-1995. In K. A. Leithwood, J. Chapmann, P. Corson, P. Hallinger, & A. Hart, *International Handbook of Educational Leadership and Administration* (Bd. 2, S. 723-781). Dordrecht: Kluwer.

Hartig, J. (2004). Methoden zur Bildung von Kompetenzstufenmodellen. In H. Moosbrugger, D. Frank, & W. Rauch, *Qualitätssicherung im Bildungswesen* (S. 74-93). Frankfurt am Main: Institut für Psychologie der Johann Wolfgang Goethe-Universität.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck, & E. Klieme, *Sprachliche Kompetenzen. Konzepte und Messung* (S. 83-99). Weinheim: Beltz.

Hartig, J., & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig, & E. Klieme, *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (S. 17-36). Berlin, Bonn: Bundesministerium für Bildung und Forschung.

Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer, *Leistung und Leistungsdiagnostik* (S. 127-143). Heidelberg: Springer.

Hartig, J., & Kühnbach, O. (2006). Schätzung von Veränderung mit "plausible values" in mehrdimensionalen Rasch-Modellen. In A. Ittel, & H. Merkens, *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft* (S. 27-44). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hartung-Beck, V. (2009). *Schulische Organisationsentwicklung und Professionalisierung. Folgen von Lernstandserhebungen an Gesamtschulen.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Heid, H. (2006). Ist die Standardisierung wünschenswerten Lernoutputs geeignet, zur Qualitätsverbesserung des Bildungswesens beizutragen? *Gymnasium Helveticum* (2), S. 19-22.

Helmke, A. (2007). *Unterrichtsqualität erfassen, bewerten, verbessern* (5. Ausg.). Seelze: Klett Kallmeyer.

Helmke, A. (2004). Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *Seminar , 2*, S. 90-112.

Helmke, A., Hosenfeld, I., & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold, & C. Griese, *Schulleitung und Schulentwicklung. Voraussetzungen, Bedingungen, Erfahrungen* (S. 119-143). Hohengehren: Schneider-Verlag.

Helmke, U. (2005). Bildungsstandards in der Unterrichtsarbeit. *Die Deutsche Schule , 97* (4), S. 448-454.

Herzog, W. (2010). Besserer Unterricht dank Bildungsstandards und Kompetenzmodellen? In A. Gehrmann, U. Hericks, & M. Lüders, *Bildungsstandards und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 37-46). Bad Heilbrunn: Klinkhardt.

Hesse, H.-G. (2008). Interkulturelle Kompetenz: Vom theoretischen Konzept über die Operationalisierung bis zum Messinstrument. In N. Jude, J. Hartig, & E. Klieme, *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte und Methoden* (S. 47-61). Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Hessischer Philologenverband. (5. April 2010). *Stellungnahme des Hessischen Philologenverbandes zu "Bildungsstandards und Inhaltsfelder - Das neue Kerncurriculum für Hessen" (Rohfassung - Stand März 2010).* Abgerufen am 22. Mai 2011 von http://www.hphv.de/media/stellungnahmen/Stellungnahme%20HPhV%20Kerncurriculum.pdf

Hessisches Kultusministerium. (2009). *Lernstandserhebungen. Aktualisierte Fassung für das Schuljahr 2009/10.* Abgerufen am 11. November 2010 von http://www.iq.hessen.de/irj/servlet/prt/portal/prtroot/slimp.CMReader/HKM_15/IQ_Internet/med/85b/85b51c29-249c-421f-012f-31e2389e4818,22222222-2222-2222-2222-222222222222

Hessisches Kultusministerium. (10.02.2011). *Schulleitungsinfo.* Wiesbaden.

Heymann, H. W. (2005a). Standards im Unterricht: Warum? Wie? Und für Wen? *PÄD-Forum: unterrichten erziehen , 33* (1), S. 23-25.

Heymann, H. W. (2005b). Tests und Unterrichtsqualität. Einführung in den Themenschwerpunkt. *Pädagogik , 57* (5), S. 6-9.

Höfer, D., Steffens, U., Diehl, G., Loleit, P., & Maier, D. (2010). *Bildungsstandards und Inhaltsfelder - Das neue Kerncurriculum für Hessen. Eine Darstellung für Lehrerinnen und Lehrer an hessischen Schulen.* Wiesbaden: Institut für Qualitätsentwicklung.

Höfer, D., Steffens, U., Diehl, G., Loleit, P., & Maier, D. (2009). *Das hessische Konzept "Bildungsstandards / Kerncurricula".* Wiesbaden: Institut für Qualitätsentwicklung.

Hofmann-Göttig, J., Eschmann, W., & Daumen, C. (2005). Und sie bewegt sich doch... Vom Umgang mit den Ergebnissen externer Evaluation aus Sicht von Bildungspolitik und Schulaufsicht. *Friedrich Jahresheft , 23*, S. 32-36.

Holtappels, H. G. (2000). Qualitätsentwicklung und Qualitätssicherung im Schulbereich. In H. Weishaupt, *Qualitätssicherung im Bildungswesen. Problemlage und aktuelle Forschungsbefunde* (S. 37-70). Erfurt: Pädagogische Hochschule Erfurt.

Horster, L., & Rolff, H.-G. (2006). *Unterrichtsentwicklung. Grundlagen einer reflektorischen Praxis* (2. Ausg.). Weinheim, Basel: Beltz.

Hosenfeld, I. (2005). Rezeption - Reflexion - Aktion. Wie lassen sich Lernstandserhebungen und Vergleichsarbeiten pädagogisch nutzen? *Friedrich Jahresheft , 23*, S. 112-114.

Hosenfeld, I., & Groß Ophoff, J. (2007). Nutzung und Nutzen von Evaluationsstudien in Schule und Unterricht. *Empirische Pädagogik , 21* (4), S. 352-367.

Hosenfeld, I., Groß Ophoff, J., & Bittins, P. (2006). *Vergleichsarbeiten und Schulentwicklung.* München: Oldenbourg.

Hovestadt, G., & Keßler, N. (2005). 16 Bundesländer. Eine Übersicht zu Bildungsstandards und Evaluation. *Friedrich Jahresheft , 23*, S. 8-10.

Husfeldt, V. (2004). Large Scale Assessments. Ihr möglicher Beitrag zur Qualitätsentwicklung von Schule und Unterricht. *Die Deutsche Schule , 96* (4), S. 500-513.

Institut für Qualitätsentwicklung. (2011). *Ergebnisse der Schulinspektion in Hessen. Berichtszeitraum 2009/2010.* Wiesbaden.

Institut für Qualitätsentwicklung. (2008). *Hessischer Referenzrahmen Schulqualität.Qualitätsbereiche, Qualitätsdimensionen und Qualitätskriterien.* Wiesbaden.

Institut für Qualitätsentwicklung im Bildungswesen. (2009a). *Handreichung VERA 8 Mathematik 2009 Testheft III.* Abgerufen am 30. November 2010 von http://www.iqb.hu-berlin.de/vera/dateien/V82009_MATTHIII.pdf

Institut für Qualitätsentwicklung im Bildungswesen. (2009b). *Kompetenzstufenmodell zu den Bildungsstandards im Fach Mathematik für den Hauptschulabschluss.* Abgerufen am 10. November 2010 von http://www.iqb.hu-berlin.de/bista/dateien/mathe_hsa.pdf

Institut für Qualitätsentwicklung im Bildungswesen. (2008). *Kompetenzstufenmodell zu den Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss.* Abgerufen am 1. Dezember 2010 von http://www.iqb.hu-berlin.de/bista/dateien/Mathe_MSA.pdf

Institut für Qualitätsentwicklung im Bildungswesen. (2009c). *Kompetenzstufenmodell zu den Bildungsstandards im Kompetenzbereich Sprechen und Zuhören - hier Zuhören - für den Mittleren Schulabschluss.* Abgerufen am 10. November 2010 von http://www.iqb.hu-berlin.de/bista/dateien/Deutsch_KSM_Hre_2.pdf

Institut für Qualitätsentwicklung im Bildungswesen. (2009d). *Kompetenzstufenmodelle zu den Bildungsstandards im Fach Englisch für den Hauptschulabschluss - Hörverstehen und Leseverstehen -.* Abgerufen am 10. November 2010 von http://www.iqb.hu-berlin.de/bista/dateien/Kompetenzen_Eng.pdf

Institut für Qualitätsentwicklung im Bildungswesen. (2009e). *Kompetenzstufenmodelle zu den Bildungsstandards im Fach Englisch für den Mittleren Schulabschluss - Hörverstehen und Leseverstehen -.* Abgerufen am 10. November 2010 von http://www.iqb.hu-berlin.de/bista/dateien/Kompetenzen_Englisch_MSA.pdf

Institut für Qualitätsentwicklung im Bildungswesen. (2007). *Perspektiven und Visionen. Tätigkeitsbericht 2005/06.* Abgerufen am 11. November 2010 von http://www.iqb.hu-berlin.de/institut/dateien/IQBJahresberich.pdf

Institut für Qualitätsentwicklung im Bildungswesen. (2010). *VERA/ Lernstandserhebungen*. Abgerufen am 11. November 2010 von http://www.iqb.hu-berlin.de/vera

Institut für Qualitätsentwicklung. (2010). *Lernstandserhebung Ergebnisbericht 8, Mathematik, Beispielschule, Klasse 8z, 2009/2010.* Wiesbaden.

Institut für Qualitätsentwicklung. (2008). *Lernstandserhebungen in Hessen. Ein Beitrag zur Schul- und Unterrichtsentwicklung.* Wiesbaden.

Institut für Qualitätsentwicklung. (2010). *Lernstandserhebungen in Hessen. Ein Beitrag zur Schul- und Unterrichtsentwicklung.* Wiesbaden.

Institut für Qualitätsentwicklung. (2009). *Statistik in der Schule.* Wiesbaden.

Institut für Qualitätsentwicklung. (2009). *Vergleichsarbeiten Mathematik 2009, Jahrgang 8, Testheft III.* Abgerufen am 1. Dezember 2010 von http://www.iqb.hu-berlin.de/vera/dateien/V82009_MATHIII_.pdf

Jude, N., & Wirth, J. (2007). Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen. In J. Hartig, & E. Klieme, *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des*

Bundesministeriums für Bildung und Forschung (S. 49-56). Berlin, Bonn: Bundesministerium für Bildung und Forschung.

Jurecka, A., & Hartig, J. (2007). Computer- und netzwerkbasiertes Assessment. In H. Johannes, & E. Klieme, *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (S. 37-48). Berlin, Bonn: Bundesministerium für Bildung und Forschung.

Keller, S., & Ruf, U. (2005). Was leisten Kompetenzmodelle? Pädagogische Konzepte für Dialogischen Unterricht am Gymnasium. *Die Deutsche Schule , 97* (4), S. 455-469.

Kiper, H. (2009). Schulentwicklung im Rahmen von Kontextsteuerung - Welche Hinweise geben (durch Evaluation und Vergleichsarbeiten gewonnene) Daten für ihre Ausrichtung? In T. Bohl, & H. Kiper, *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 13-28). Bad Heilbrunn: Klinkhardt.

Klemm, K. (2000). Large scale assessments in einem modernisierten Bildungssystem. *Die Deutsche Schule , 92* (3), S. 329-338.

Klieme, E. (2003). Bildungsstandards. Ihr Beitrag zur Qualitätsentwicklung im Schulsystem. *Die Deutsche Schule , 95* (1), S. 10-16.

Klieme, E. (2004). Was sind Kompetenzen und wie lassen sie sich messen? *Pädagogik , 56* (6), S. 10-13.

Klieme, E. (2005). Zur Bedeutung von Evaluation für die Schulentwicklung. In K. Maag Merki, A. Sandmeier, P. Schuler, H. Fend, E. Klieme, H. Faulstich-Wieland, et al., *Schule wohin? Schulentwicklung und Qualitätsmanagement im 21. Jahrhundert* (S. 40-61). Zürich: Universität Zürich.

Klieme, E., & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin, & H.-H. Krüger, *Kompetenzdiagnostik* (S. 11-29). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik , 52* (6), S. 876-903.

Klieme, E., Artelt, C., & Stanat, P. (2001). Fächerübergreifende Kompetenzen. Konzepte und Indikatoren. In F. E. Weinert, *Leistungsmessungen in Schulen* (S. 203-218). Weinheim, Basel: Beltz.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (2007). *Zur Entwicklung nationaler Bildungsstandards. Expertise* (unveränderte Ausg.). Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Klinger, U. (2005). Mit Bildungsstandards Unterrichts- und Schulqualität entwickeln. Eine Curriculumwerkstatt für Fachkonferenzen, Steuergruppen und Schulleitungen. *Friedrich Jahresheft , 23*, S. 130-143.

Klippert, H. (1997). Schule entwickeln - Unterricht neu gestalten. *Pädagogik , 49* (2), S. 12-17.

Klug, C., & Reh, S. (2000). Was fangen die Schulen mit den Ergebnissen an? Die Hamburger Leistungsvergleichsstudie aus der Sicht "beforschter" Schulen. *Pädagogik , 52* (12), S. 16-21.

Koch, U., Groß Ophoff, J., Hosenfeld, I., & Helmke, A. (2006). Qualitätssicherung: Von der Evaluation zur Schul- und Unterrichtsentwicklung. Ergebnisse der Lehrerbefragungen zur Auseinandersetzung mit den VERA-Rückmeldungen. In F. Eder, A. Gasteiger, & F. Hofmann, *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 187-199). Münster, New York, München, et al.: Waxmann.

Kohler, B. (2005). *Rezeption internationaler Schulleistungsstudien. Wie gehen Lehrkräfte, Eltern und die Schulaufsicht mit Ergebnissen schulischer Evaluationsstudien um?* Münster, New York, München, et al.: Waxmann.

Kohler, B. (2009). Umgang von Lehrer/innen, Eltern und Schulaufsicht mit Ergebnissen internationaler Schulleistungsstudien. In T. Bohl, & H. Kiper, *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 81-96). Bad Heilbrunn: Klinkhardt.

Kohler, B., & Schrader, F.-W. (2004). Ergebnisrückmeldung und Rezeption: Von der externen Evaluation zur Entwicklung von Schule und Unterricht. *Empirische Pädagogik , 18* (1), S. 3-17.

Köller, O. (2008a). Bildungsstandards - Verfahren und Kriterien bei der Entwicklung von Messinstrumenten. *Zeitschrift für Pädagogik , 54* (2), S. 163-173.

Köller, O. (2010). Bildungsstandards. In R. Tippelt, & B. Schmidt, *Handbuch Bildungsforschung* (S. 529-548). Wiesbaden: VS Verlag für Sozialwissenschaften.

Köller, O. (2007). Standards und Qualitätssicherung zur Outputsteuerung im System und in der Einzelinstitution. In J. van Buer, & C. Wagner, *Qualität von Schule. Ein kritisches Handbuch* (S. 93-102). Frankfurt am Main, Berlin, Bern, et al.: Lang.

Köller, O. (2008b). *Zum Verhältnis von Kompetenzstufen, Normierung der Bildungsstandards und standardisierten Lernstandserhebungen. Präsentation auf der 8. EMSE-*

Fachtagung am 05./06. Juni 2008 in Wiesbaden. Abgerufen am 1. Dezember 2010 von http://www.emse-netzwerk.de/uploads/Main/EMSE08_Koeller_Verhaeltnis_von_Kompetenzstufen_Bildungsstandards_und_LSE.pdf

Korngiebel, J. (2009). *Kompetenztests in der Sekundarstufe I als Überprüfungsinstrument der Bildungsstandards.* Marburg: Tectum.

Kreitz, R. (2010). Was ist es, was Kompetenztests messen? In A. Gehrmann, U. Hericks, & M. Lüders, *Bildungsstandards und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 55-69). Bad Heilbrunn: Klinkhardt.

Kuckartz, U. (2010). *Einführung in die computergestützte Analyse qualitativer Daten* (3. akt. Aufl. Ausg.). Wiesbaden: VS Verlag.

Kuckartz, U., Dresing, T., Rädiker, S., & Stefer, C. (2008). *Qualitative Evaluation. Der Einstieg in die Praxis* (2. akt. Aufl. Ausg.). Wiesbaden: VS Verlag.

Kühle, B., & Peek, R. (2007). Lernstandserhebungen in Nordrhein-Westfalen. Evaluationsbefunde zur Rezeption und zum Umgang mit Ergebnisrückmeldungen in Schulen. *Empirische Pädagogik , 21* (4), S. 428-447.

Kultusministerkonferenz. (2005a). *Bildungsstandards der Kultusministerkonferenz. Erläuterungen zur Konzeption und Entwicklung.* München, Neuwied: Luchterhand.

Kultusministerkonferenz. (2005b). *Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4). Beschluss der Kultusministerkonferenz vom 15.10.2004.* München: Luchterhand.

Kultusministerkonferenz. (2010). *Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung. Beschluss der Kultusministerkonferenz vom 10.12.2009.* Köln: Carl Link.

Kultusministerkonferenz. (1997). Pressemitteilung zur 280. Plenarsitzung der Kultusministerkonferenz am 23./24. Oktober 1997 in Konstanz.

Kultusministerkonferenz. (2001). Pressemitteilung zur 296. Plenarsitzung der Kultusministerkonferenz am 05./06. Dezember 2001 in Bonn.

Kultusministerkonferenz. (2002a). Pressemitteilung zur 298. Plenarsitzung der Kultusministerkonferenz am 23./24. Mai 2002 in Eisenach.

Kultusministerkonferenz. (2002b). Pressemitteilung zur 299. Plenarsitzung der Kultusministerkonferenz am 17./18. Oktober 2002 in Würzburg.

Kultusministerkonferenz. (2006). Pressemitteilung zur 314. Plenarsitzung der Kultusministerkonferenz am 01./02. Juni 2006 in Plön.

Kultusministerkonferenz. (2007). Pressemitteilung zur 319. Plenarsitzung der Kultusministerkonferenz am 17./18. Oktober 2007 in Bonn.

Kultusministerkonferenz. (2004a). Vereinbarung über Bildungsstandards für den Hauptschulabschluss (Jahrgangsstufe 9). Beschluss der Kultusministerkonferenz vom 15.10.2004.

Kultusministerkonferenz. (2003). Vereinbarung über Bildungsstandards für den Mittleren Bildungsabschluss (Jahrgangsstufe 10). Beschluss der Kultusministerkonferenz vom 04.12.2003.

Kultusministerkonferenz. (2004b). Vereinbarung über Bildungsstandards für den Mittleren Schulabschluss (Jahrgangsstufe 10) in den Fächern Biologie, Chemie, Physik. Beschluss der Kultusministerkonferenz vom 16.12.2004.

Kultusministerkonferenz. (2004c). Vereinbarung über Bildungsstandards für den Primarbereich (Jahrgangsstufe 4). Beschluss der Kultusministerkonferenz vom 15.10.2004.

Kuper, H. (2008). Entscheiden und Kommunizieren. Eine Skizze zum Wandel schulischer Leitungs- und Partizipationsstrukturen und den Konsequenzen für die Lehrerprofessionalität. In W. Helsper, S. Busse, M. Hummrich, & R.-T. Kramer, *Pädagogische Professionalität in Organisationen. Neue Verhältnisbestimmungen am Beispiel der Schule* (S. 149-162). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kuper, H., & Hartung, V. (2007). Überzeugungen zur Verwendung des Wissens aus Lernstandserhebungen. EIne professionstheoretische Analyse. *Zeitschrift für Erziehungswissenschaft , 10* (2), S. 214-229.

Lange, B. (2005). Bildungsstandards und Unterrichtsplanung - Konsequenzen für didaktisches Denken und Planen. *Lehren und lernen , 31* (1), S. 3-10.

Lankes, E.-M. (2006). Mit Bildungsstandards arbeiten - kompetenzorientiert unterrichten. *Grundschule , 38* (5), S. 21-23.

Lersch, R. (2007). Unterricht und Kompetenzerwerb. In 30 Schritten von der Theorie zur Praxis kompetenzfördernden Unterrichts. *Die Deutsche Schule , 99* (4), S. 434-446.

Lersch, R. (2006). Unterricht zwischen Standardisierung und individueller Förderung. Überlegungen zu einer neuen Lernkultur angesichts der bevorstehenden Einführung von Bildungsstandards. *Die Deutsche Schule , 98* (1), S. 28-40.

Lersch, R. (2010). *Wie unterrichtet man Kompetenzen? Didaktik und Praxis kompetenzfördernden Unterrichts.* Wiesbaden: Institut für Qualitätsentwicklung.

Leutner, D., Fleischer, J., Spoden, C., & Wirth, J. (2007). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik. In M. Prenzel, I. Gogolin, &

H.-H. Krüger, *Kompetenzdiagnostik* (S. 149-167). Wiesbaden: VS Verlag für Sozialwissenschaften.

Levin, A. (2009). *Qualitätsprobleme mathematischer Vergleichsarbeiten. Erfassung mathematischer Kompetenzen und psychometrische Modellierung einer landesweiten Prüfungsarbeit in Klassenstufe 10.* Münster, New York, München, et al.: Waxmann.

Lienert, G. A. (1969). *Testaufbau und Testanalyse.* Weinheim, Basel: Beltz.

Lorenz, J. H. (2005). Zentale Lernstandsmessungen in der Primarstufe - Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern. *Zentralblatt für Didaktik der Mathematik , 37* (4), S. 317-323.

Maag Merki, K. (2010). Theoretische und empirische Analyse der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter, & K. Maag Merki, *Handbuch Neue Steuerung im Schulsystem* (S. 145-169). Wiesbaden: VS Verlag für Sozialwissenschaften.

Maag Merki, K., & Steinert, B. (2006). Die Prozessstruktur von teilautonomen Schulen und ihre Effektivität für die Herstellung optimaler Lernkontexte für schulische Bildungsprozesse. *Schweizerische Zeitschrift für Bildungswissenschaften , 28*, S. 103-122.

Maier, U. (2010a). Effekte testbasierter Rechenschaftslegung auf Schule und Unterricht. Ist die internationale Befundlage auf Vergleichsarbeiten im deutschsprachigen Raum übertragbar? *Zeitschrift für Pädagogik , 56* (1), S. 112-128.

Maier, U. (2010b). Formative und summative Aspekte testbasierter Schulreformen. In A. Gehrmann, U. Hericks, & M. Lüders, *Bildungsstandards und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 47-54). Bad Heilbrunn: Klinkhardt.

Maier, U. (2007). *Lehrereinschätzungen zu zentralen Tests und Leistungsrückmeldungen. Ein Vergleich zwischen Baden-Württemberg und Thüringen. Präsentation auf der 7. EMSE-Fachtagung am 6./7. Dezember 2007 in Mainz.* Abgerufen am 7. Januar 2011 von http://www.emse-netzwerk.de/uploads/Main/EMSE07_Maier_Lehrereinschaetzungen.pdf

Maier, U. (2009). Professionelle Nutzung von Vergleichsarbeiten? - Ergebnisse einer qualitativen Interviewstudie mit Lehrkräften in Baden-Württemberg. In T. Bohl, & H. Kiper, *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 131-144). Bad Heilbrunn: Klinkhardt.

Maier, U. (2008a). Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. *Zeitschrift für Pädagogik , 54* (1), S. 95-117.

Maier, U. (2008b). Vergleichsarbeiten im Vergleich - Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen. *Zeitschrift für Erziehungswissenschaft , 11* (3), S. 453-474.

Maier, U. (2008c). Was lernen Schulen aus zentralen Tests? Was sollten Bildungspolitiker lernen, wenn sie testen lassen? *Die Deutsche Schule , 100* (1), S. 66-72.

Maier, U., & Rauin, U. (2006). Vergleichsarbeiten - Hilfe zur Unterrichtsentwicklung? Zentrale Lernstandserhebungen aus Sicht baden-württembergischer Lehrkräfte. *Die Deutsche Schule , 98* (4), S. 403-421.

Malik, F. (2001). *Führen, Leisten, Leben - Wirksames Management für eine neue Zeit.* München: Heyne.

Markstahler, J., Schwarz, A., & Steffens, U. (2004). PISA 2000 in Hessen. Schulrückmeldung braucht Schulberatung. *Schulverwaltung. Ausgabe Hessen, Rheinland-Pfalz und Saarland , 8* (7-8), S. 200-202.

Mayer, H. O. (2006). *Interview und schriftliche Befragung. Entwicklung, Durchführung und Auswertung* (3. überarb. Aufl. Ausg.). München, Wien: Oldenbourg Verlag.

Mayring, P. (2007). *Qualitative Inhaltsanalyse. Grundlagen und Techniken.* Weinheim: Deutscher Studien Verlag.

Meetz, F. (2007). *Personalentwicklung als Element von Schulentwicklung. Bestandsaufnahme und Perspektiven.* Bad Heilbrunn: Klinkhardt.

Meinel, P., & Sachse, M. (2007). Landesinstitut und Schulentwicklung - Unterstützungssystem zwischen Bildungsadministration und Schulpraxis. In J. van Buer, & C. Wagner, *Qualität von Schule. Ein kritisches Handbuch* (S. 259-272). Frankfurt am Main, Berlin, Bern, et al.: Lang.

Mentzel, W. (1997). *Unternehmenssicherung durch Personalentwicklung. Mitarbeiter motivieren, fördern und weiterbilden* (7. Ausg.). Freiburg im Breisgau: Haufe.

Meyerhöfer, W. (2005). *Tests im Test: Das Beispiel PISA.* Opladen: Budrich.

Moser, U., & Keller, F. (2002). *Wie beurteilen Lehrpersonen die Rückmeldungen von Ergebnissen bei Leistungsevaluationen? Eine Befragung von Lehrpersonen im Anschluss an die Evaluation der 3. Primarschulklassen.* Zürich: Bildungsdirektion des Kantons Zürich.

Münch, J. (2004). Notwendigkeiten, Möglichkeiten und Grenzen von Führungshandeln in selbstständigen Schulen. In R. Arnold, & C. Griese, *Schulleitung und Schulentwicklung. Voraussetzungen, Bedingungen, Erfahrungen* (S. 25-40). Hohengehren: Schneider-Verlag.

Nachtigall, C. (2009). *Landesbericht Thüringer Kompetenztests 2009.* Abgerufen am 7. Januar 2011 von http://www.kompetenztest.de/download/kt2009/KT2009-Landesbericht.pdf

Nachtigall, C. (2010). *Landesbericht Thüringer Kompetenztests 2010.* Abgerufen am 7. Januar 2011 von http://www.kompetenztest.de/download/kt2010/KT2010_Landesbericht.pdf

Nachtigall, C., & Jantowski, A. (2007). Die Thüringer Kompetenztests unter besonderer Berücksichtigung der Evaluationsergebnisse zum Rezeptionsverhalten. *Empirische Pädagogik , 21* (4), S. 401-410.

Nachtigall, C., & Kröhne, U. (2006). Methodische Anforderungen an schulische Leistungsmessung - auf dem Weg zu fairen Vergleichen. In H. Kuper, & J. Schneewind, *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 59-74). Münster, New York, München, et al.: Waxmann.

Netzwerk Empiriegestützte Schulentwicklung. (2008). *Nutzung und Nutzen von Schulrückmeldungen im Rahmen standardisierter Lernstandserhebungen/ Vergleichsarbeiten. Zweites Positionspapier des EMSE-Netzwerkes - verabschiedet auf der 9. EMSE-Fachtagung am 16.-17. Dezember 2008 in Nürnberg.* Abgerufen am 11. November 2010 von http://www.emse-netzwerk.de/uploads/Main/EMSE_Positionsp2_Rueckmeldungen.pdf

Netzwerk Empiriegestützte Schulentwicklung. (2006). *Positionspapier des Netzwerks Empiriegestützte Schulentwicklung (EMSE) zu: Zentrale standardisierte Lernstandserhebungen. 5. EMSE-Fachtagung in Berlin am 08.12.2006.* Abgerufen am 11. 11 2010 von http://www.emse-netzwerk.de/uploads/Main/EMSE_Positionsp1_Lernstandserh.pdf

Neubrand, M. (2004). *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Neumann, D., Karius, I., Robitzsch, A., Behrens, U., Krelle, M., & Böhme, K. (kein Datum). *Der Pozess der Aufgabenentwicklung am Beispiel einer Aufgabe.* Abgerufen am 1. Dezember 2010 von Institut für Qualitätsentwicklung im Bildungswesen: http://www.iqb.hu-berlin.de/arbbereiche/testentw/projekte/dateien/Aufgabenentwick.pdf

Niedermair, G., & Bachmann, H. (2002). Personalentwicklung. In F. Eder, & P. Posch, *Qualitätsentwicklung und Qualitätssicherung im österreichischen Schulwesen* (S. 101-108). Innsbruck, Wien, München, et al.: Studien-Verlag.

Oelkers, J. (2009). *Einige Gelingensbedingungen für kompetenzorientierten Unterricht. Vortrag auf der 10. EMSE-Fachtagung am 19. Juni 2009 in Dresden.* Abgerufen am 3. Januar 2011 von http://www.emse-netzwerk.de/uploads/Main/EMSE10_Material_Oelker.pdf

Orth, G. (2001). Vergleichsarbeiten. In H.-G. Rolff, & H.-J. Schmidt, *Brennpunkt Schulleitung und Schulaufsicht. Konzepte und Anregungen für die Praxis* (S. 203-223). Neuwied: Luchterhand.

Peek, R. (2009). Dateninduzierte Schulentwicklung. In H. Buchen, & H.-G. Rolff, *Professionswissen Schulleitung* (2. Ausg., S. 1343-1367). Weinheim, Basel: Beltz.

Peek, R. (2004a). Klassenbezogene Rückmeldungen aus Schulleistungsstudien und ihre Rezeption in beteiligten Schulen im Land Brandenburg (Projekt QuaSUM 2). In R. Peek, & I. Nilshon, *Schulrückmeldungen im Rahmen von Schulleistungsstudien am Beispiel des QuaSUM-Projektes. Zwei Untersuchungen zur Wirksamkeit* (S. 9-107). Potsdam: Ministerium für Bildung, Jugend und Sport des Landes Brandenburg.

Peek, R. (2004b). Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik (QuaSUM) - Klassenbezogene Ergebnisrückmeldungen und ihre Rezeption in Brandenburger Schulen. *Empirische Pädagogik , 18* (1), S. 82-114.

Peek, R., & Dobbelstein, P. (2006). Benchmarks als Input für die Schulentwicklung - das Beispiel der Lernstandserhebungen in Nordrhein-Westfalen. In H. Kuper, & J. Schneewind, *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 41-58). Münster, New York, München, et al.: Waxmann.

Peek, R., Pallack, A., Dobbelstein, P., Fleischer, J., & Leutner, D. (2006). Lernstandserhebungen 2004 in Nordrhein-Westfalen - zentrale Testergebnisse und Perspektiven für die Schul- und Unterrichtsentwicklung. In E. Ferdinand, A. Gasteiger, & F. Hofmann, *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 219-233). Münster, New York, München, et al.: Waxmann.

Posch, P. (2009). Zur schulpraktischen Nutzung von Daten: Konzepte, Strategien, Erfahrungen. *Die Deutsche Schule , 101* (2), S. 119-135.

Posch, P., & Altrichter, H. (1997). *Möglichkeiten und Grenzen der Qualitätsevaluation und Qualitätsentwicklung im Schulwesen.* Innsbruck: Studien-Verlag.

Ramm, G. (2006). *Lernstand 6 (VERA 6). Präsentation auf der 4. EMSE-Fachtagung im Juni 2006 in Hamburg.* Abgerufen am 30. November 2010 von http://www.emse-netzwerk.de/uploads/Main/EMSE04.zip

Regenbrecht, A. (2005). Sichern Bildungsstandards die Bildungsaufgabe der Schule? *PÄD-Forum: unterrichten erziehen , 33* (1), S. 16-22.

Reh, S. (2008). "Reflexivität der Organisation" und Bekenntnis. Perspektiven der Lehrerkooperation. In W. Helsper, S. Busse, M. Hummrich, & R.-T. Kramer, *Pädagogische Professionalität in Organisationen. Neue Verhältnisbestimmungen am Beispiel der Schule* (S. 163-183). Wiesbaden: VS Verlag für Sozialwissenschaften.

Reinders, H. (2008). Erfassung sozialer und selbstregulatorischer Kompetenzen bei Kindern und Jugendlichen - Forschungsstand. In N. Jude, J. Hartig, & E. Klieme, *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte und Methoden* (S. 27-45). Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Rolff, H.-G. (1986). Selbsterneuerung durch Organisationsentwicklung. *Schulmanagement , 17* (3), S. 12-15.

Rolff, H.-G. (2007a). Steuergruppen als Basis von Schulentwicklung. In N. Berkemeyer, & G. G. Holtappels, *Schulische Steuergruppen und Change Management* (S. 41-60). Weinheim, München: Juventa.

Rolff, H.-G. (2007b). *Studien zu einer Theorie der Schulentwicklung.* Weinheim, Basel: Beltz.

Rolff, H.-G. (1993). *Wandel durch Selbstorganisation. Theoretische Grundlagen und praktische Hinweise für eine bessere Schule* (2. Ausg.). Weinheim, München: Juventa.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation - A systematic approach* (7. Ausg.). Thousand Oaks, London, New Dehli: Sage.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion.* Bern, Göttingen, Toronto, et al.: Huber.

Ryan, R. M., & Sapp, A. (2005). Zum Einfluss testbasierter Reformen: High Stakes Testing (HST). Motivation und Leistung aus Sicht der Selbstbestimmungstheorie. *Unterrichtswissenschaft , 33* (2), S. 143-159.

Schirp, H. (2006a). Zentrale Leistungstests und ihre Auswirkungen - ein Blick in die USA. *Schulverwaltung. Hessen, Rheinland-Pfalz , 11* (10), S. 267-270.

Schirp, H. (2006b). Zentrale quantitative Leistungsmessungen und qualitative Schulentwicklung. Die Wirkungen von High Stakes Tests in den USA. *Die Deutsche Schule , 98* (4), S. 422-435.

Schley, W. (2004). Evaluation als Intervention durch Feedback. *Journal für Schulentwicklung , 8* (1), S. 16-25.

Schneewind, J. (2007a). Erfahrungen mit Ergebnisrückmeldungen im Projekt BeLesen - Ergebnisse der Interviewstudie. *Empirische Pädagogik , 21* (4), S. 368-382.

Schneewind, J. (2006). Rückmeldungen als Motivator für die Teilnahme an Schulleistungsstudien? Die Rezeptionsstudie von BeLesen. In H. Kuper, & J. Schneewind, *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 107-126). Münster, New York, München, et al.: Waxmann.

Schneewind, J. (2007b). *Wie Lehrkräfte mit Ergebnisrückmeldungen aus Schulleistungsstudien umgehen. Ergebnisse aus Befragungen von Berliner Grundschullehrerinnen.* Diss., Berlin: Freie Univ.

Schneewind, J., & Kuper, H. (2009). Rückmeldeformate und Verwendungsmöglichkeiten der Ergebnisse aus zentralen Lernstandserhebungen. In T. Bohl, & H. Kiper, *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 113-129). Bad Heilbrunn: Klinkhardt.

Schneewind, J., Merkens, H., & Kuper, H. (2005). Erprobung eines Rückmeldeformats an Berliner Grundschulen. In H. Döbert, & H.-W. Fuchs, *Leistungsmessungen und Innovationsstrategien in Schulsystemen. Ein internationaler Vergleich* (S. 70-94). Münster, New York, München, et al.: Waxmann.

Schrader, F.-W., & Helmke, A. (2004). Von der Evaluation zur Innovation? Die Rezeptionsstudie WALZER: Ergebnisse der Lehrerbefragung. *Empirische Pädagogik , 18* (1), S. 140-161.

Schweitzer, K. (2007). *Der Schwierigkeitsgrad von Textverstehensaufgaben. Ein Beitrag zur Differenzierung und Präzisierung von Aufgabenbeschreibungen.* Frankfurt am Main: Peter Lang.

Schweizer, K., & Klieme, E. (2005). Kompetenzstufen der Lehrerkooperation. Ein empirisches Beispiel für das Latent-Growth-Curve-Modell. *Psychologie in Erziehung und Unterricht , 52* (1), S. 66-79.

Schwippert, K. (2004). Leistungsrückmeldungen an Grundschulen im Rahmen der Internationalen Grundschul-Lese-Untersuchung (IGLU). *Empirische Pädagogik , 18* (1), S. 62-81.

Schwippert, K. (2005a). Tests. Oder: Wie man Äpfel mit Birnen vergleicht. *Friedrich Jahresheft , 23*, S. 15-17.

Schwippert, K. (2005b). Vergleichende Lernstandsuntersuchungen, Bildungsstandards und die Steuerung von schulischen Bildungsprozessen. *Berufs- und Wirtschaftspädagogik - online* (8), S. 1-14.

Speck-Hamdan, A. (2007). Entwicklung von Unterrichtsqualität durch Standards? In K. Möller, P. Hanke, C. Beinbrech, A. K. Hein, T. Kleickmann, & R. Schages, *Qualität von Grundschulunterricht entwickeln, erfassen und bewerten* (S. 91-94). Wiesbaden: VS Verlag für Sozialwissenschaften.

Staatsinstitut für Schulqualität und Bildungsforschung München. (2006). *Kompetenz... mehr als nur Wissen! Informationsblatt.* Abgerufen am 10. November 2010 von http://www.kompas.bayern.de/userfiles/infokompetenz.pdf

Stähling, R. (2005). Qualitätsentwicklung statt Vergleichsarbeiten. Zu einem unfruchtbaren Verhältnis von Forschung und Schule. *Die Deutsche Schule , 97* (2), S. 211-221.

Steffens, U. (2009). *Einleitung in die 10. EMSE-Fachtagung am 18./19. Juni 2009 in Dresden.* Abgerufen am 10. November 2010 von http://www.emse-netzwerk.de/uploads/Main/EMSE10_Einleitu.pdf

Steffens, U. (2007). Schulqualitätsdiskussion in Deutschland - Ihre Entwicklung im Überblick. In J. van Buer, & C. Wagner, *Qualität von Schule. Ein kritisches Handbuch* (S. 21-51). Frankfurt am Main, Berlin, Bern, et al.: Lang.

Steinert, B., Klieme, E., Maag Merki, K., Döbrich, P., Halbheer, U., & Kunz, A. (2006). Lehrerkooperation in der Schule: Konzeption, Erfassung, Ergebnisse. *Zeitschrift für Pädagogik , 52* (2), S. 185-204.

Steinweg, A. S. (2007). Mathematik auf dem Prüfstand. *Die Grundschulzeitschrift , 21* (207), S. 4-9.

Stern, E., & Hardy, I. (2002). Schulleistungen im Bereich der mathematischen Bildung. In F. E. Weinert, *Leistungsmessungen in Schulen* (2. Ausg., S. 153-168). Weinheim, Basel: Beltz.

Terhart, E. (2010). Personalauswahl, Personaleinsatz und Personalentwicklung an Schulen. In H. Altrichter, & K. Maag Merki, *Handbuch Neue Steuerung im Schulsystem* (S. 255-275). Wiesbaden: VS Verlag für Sozialwissenschaften.

Thies, E. (2005). Die Entwicklung von Bildungsstandards als Länder übergreifendes bildungspolitisches Programm. In J. Rekus, *Bildungsstandards, Kerncurricula und die Aufgabe der Schule* (S. 8-16). Münster: Aschendorff.

Thüringer Kultusministerium. (2009). *Lehrermanual II. Kompetenztest Deutsch - Teile 1 und 2 in der Klassenstufe 6 im Schuljahr 2008/2009.* Abgerufen am 1. Dezember 2010 von http://www.thueringen.de/imperia/md/content/tkm/informationen/kompetenztests/2009/d6_hinweise.pdf

Traub, A. (2008). Erfassung sozialer Kompetenzen im Kinderpanel. In N. Jude, J. Hartig, & E. Klieme, *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte*

und Methoden (S. 63-72). Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Tresch, S. (2007). *Potenzial Leistungstest. Wie Lehrerinnen und Lehrer Ergebnisrückmeldungen zur Sicherung und Steigerung ihrer Unterrichtsqualität nutzen.* Bern: h.e.p.-Verlag.

Uhle, R. (2007). Bildungsstandards. In A. Henschel, & R. Krüger, *Jugendhilfe und Schule. Handbuch für eine gelingende Kooperation* (S. 39-54). Wiesbaden: VS Verlag für Sozialwissenschaften.

Ulich, K. (1996). *Beruf: Lehrer/in. Arbeitsbelastung, Beziehungskonflikte, Zufriedenheit.* Weinheim, Basel: Beltz.

van Ackeren, I. (2003). *Evaluation, Rückmeldung und Schulentwicklung. Erfahrungen mit zentralen Tests, Prüfungen und Inspektionen in England, Frankreich und den Niederlanden.* Münster, New York, München, et al. : Waxmann.

van Ackeren, I., & Bellenberg, G. (2004). Parallelarbeiten, Vergleichsarbeiten und Zentrale Abschlussprüfungen. Bestandsaufnahme und Perspektiven. *Jahrbuch der Schulentwicklung , 13*, S. 125-159.

van den Heuvel-Panhuizen, M. (2007). "Nicht nur die Antworten zählen". Wege zu einer veränderten Testkultur. *Die Grundschulzeitschrift , 21* (207), S. 10-12.

van den Heuvel-Panhuizen, M. (2006). Wie groß muss der Teppich sein? Erfahrungen mit Bildungsstandards und Leistungsmessung aus den Niederlanden. *Grundschule , 38* (5), S. 14-17.

van Weeren, J. (2007). Wem nutzen Outputmessungen? Eine kritische Analyse ihrer Wirksamkeit und Nebeneffekte aus niederländischer Perspektive. *Die Deutsche Schule , 99* (2), S. 210-223.

Visscher, A. J., & Coe, R. (2003). School performance feedback systems. Conceptualisation, Analysis and Reflection. *School Effectiveness and School Improvement , 14* (3), S. 321-349.

Vortmann, H. (2005). Bildungsstandards - Kerncurricula - Vergleichsarbeiten. Ein Bundesländervergleich. In J. Rekus, *Bildungsstandards, Kerncurricula und die Aufgabe der Schule* (S. 108-135). Münster: Aschendorff.

Watermann, R., Stanat, P., Kunter, M., Klieme, E., & Baumert, J. (2003). Schulrückmeldungen im Rahmen von Schulleistungsuntersuchungen: Das Disseminationskonzept von PISA-2000. *Zeitschrift für Pädagogik , 49* (1), S. 92-111.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert, *Leistungsmessungen in Schulen* (S. 17-31). Weinheim, Basel: Beltz.

Weinert, F. E., Schrader, F.-W., & Helmke, A. (1990). Educational expertise: Closing the gap between educational research and classroom practice. *School Psychology International , 11*, S. 163-180.

Wolff, E. (2009). *Kompetenzorientierte Unterrichtsentwicklung in Sachsen - Konzepte und Modelle. Präsentation auf der 10. EMSE-Fachtagung am 18./19. Juni 2009 in Dresden.* Abgerufen am 10. November 2010 von http://www.emse-netzwerk.de/uploads/Main/EMSE10_Wolff.pdf

Zeitler, S., Köller, O., & Tesch, B. (2010). Bildungsstandards und ihre Implikationen für Qualitätssicherung und Qualitätsentwicklung. In A. Gehrmann, U. Hericks, & M. Lüders, *Bildungsstandards und Kompetenzmodelle. Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 23-36). Bad Heilbrunn: Klinkhardt.

Ziener, G. (2006). *Bildungsstandards in der Praxis. Kompetenzorientiert unterrichten.* Seelze-Velber: Kallmeyer Klett.