6 Das Messmodell. Messverfahren zur Analyse fallbasierter Diagnosekompetenzen

Hendrik Baumbach

Im Zuge des DFG-Schwerpunktprogramms "Kompetenzdiagnostik" wurden seit 2006 für das gleichnamige Arbeitsgebiet erstens theoretisch fundierte Kompetenzmodelle und zweitens deren Umsetzung in empirisch prüfbare Modellen und Messverfahren gefordert (vgl. Klieme & Hartig 2008: 26). Anlass für die nachfolgende Formulierung eines neuartigen Modellierungs- und Messansatzes ist der Einsatz von Fallvignetten in Kompetenztests verschiedener Berufsfelder wie der Lehrerbildung, der Pädagogik und der Psychologie. Die offene Komplexität solcher Fallvignetten hat die bisher zur Auswertung von Fragebögen und Aufgabenlösungen (z.B. Schulleistungsstudien wie TIMSS, PISA) verwendeten statistischen Verfahren einer neuen Anforderungssituation ausgesetzt. Diese Neumodellierung folgt den Prämissen des vorgestellten Kompetenzmodells (vgl. Dirks 2012b, Kap. II.5) zur Struktur und Formulierung fallbezogener Items, zu den Prinzipien der Auswertung sowie zu den Anforderungen an die Maßgröße, mit dem Ziel, die in Textform von Fallanalysen erhobenen Testdaten zum Vergleich zu quantifizieren. Die anstehenden Ausführungen sind projektbezogen entwickelt worden und verstehen sich explizit als Propositionen für weitere Forschungsvorhaben, die sich mit einem ähnlichen Testformat der Kompetenzmessung beschäftigen. Qualitative Methoden werden hierbei kaum berücksichtigt, werden aber freilich nicht als nachrangig betrachtet; vielmehr soll diese Modellierung mit einem Ansatz zur Quantifizierung der Testdaten auch qualitative Auswertungen ergänzen oder zur Bildung von Hypothesen beitragen.

6.1 Modellannahmen

Die Messung diagnostischer Kompetenz anhand fallanalytischer Arbeit bedeutet für den Modellierungsprozess eine Vielzahl von Herausforderungen. Diese gehen über die inzwischen in der Kompetenzforschung standardisiert eingesetzten Testverfahren, zu welchen zuerst die variationsreichen Instrumente der probabilistischen Testtheorie – die Modelle der Item-Response-Theorie (IRT) – zählen, hinaus. Dabei unterscheidet sich das im Forschungsprojekt eingesetzte Testformat, d.h. die Bearbeitung einer Fallvignette unter einer offenen Frag-

stellung und einem offenen Antwortformat, von enger definierten Item- und Fragebogenkonstruktionen. Zielsetzung des ersten Abschnittes ist die Zusammenfassung aller Modellannahmen und die Einführung der für die Messung notwendigen Kenngrößen zur Entwicklung eines eher der klassischen Testtheorie (KTT) zugehörigen Messverfahrens.

6.1.1 Testtheoretische Rahmung

Wenngleich sich die Messung pädagogisch soziologischer Diagnosekompetenz nicht mehr auf den Begriff der Lösungswahrscheinlichkeit der verschiedenen Items beziehen soll, werden verschiedene Grundannahmen der IRT auch in der Neumodellierung verwendet und in adaptierter Form in das Messverfahren implementiert. Als Begründung für dieses Vorgehen mag an dieser Stelle ausreichen, dass die Entwicklung von Kompetenzmodellen in der pädagogisch-psychologischen Forschung vermehrt auf den mathematisch-statistischen Ansatz der IRT zurückgegriffen hat und demfolgend in ihm auch Vorteile gegenüber den (älteren) Methoden der KTT erkannt haben muss. Urform der probabilistischen Modelle ist das eindimensionale Rasch-Modell, aus welchem kontinuierlich für verschiedene Anforderungen weitere, z.T. restriktivere Modelle entstanden sind (vgl. Rasch 1960, DeMars 2010 etc.). Um die benötigten Grundannahmen dieser Modellierung dem Leser[1] in einfacher Form zugänglich zu machen, wurde für die folgenden Ausführungen die prägnante Darstellung nach Strobl (2010: 8f.) gewählt.

Auszugehen ist von einer zweidimensionalen Datenmatrix, deren Zeilen den n untersuchten Probandenund deren Spalten den m verwendeten Items (z.B. Fragen oder Aufgaben) entsprechen. Eine Quantifizierung wird nun durch eine Modellgleichung (1) erreicht, die die Wahrscheinlichkeit, ein Item zu lösen, in Abhängigkeit von den Itemschwierigkeiten β_j und der Fähigkeit der Probanden θ_i angibt. Beide Größen werden aus den Spalten- bzw. Zeilensummen der Datenmatrix geschätzt und sind auf der gleichen Skala abgetragen:

$$P(X_{ij} = 1 \mid \theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}. \tag{1}$$

X_{ij} stellt hier eine Zufallsvariable dar, die im Rahmen der Datenerhebung einer Matrix mit m Spalten und n Zeilen entspricht, wobei x_{ij} als Eintrag dieser Datenmatrix in der i-ten Spalte und j-ten Zeile zu lesen ist. Wird diese Modellierung in die Kompetenzmessung übertragen, dann wird die Wahrscheinlichkeit auf der linken Seite der Modellgleichung (1) als *relatives Leistungspotential einer Person* interpretiert. Dieses Vorgehen soll gerade zum Ausdruck bringen, dass z.B. Items niedriger Schwierigkeit bei hoher Personen-Fähigkeit in der Realität nicht sicher, sondern nur mit großer Wahrscheinlichkeit gelöst werden. Auf diese Weise kann bei bekannter Personen-Fähigkeit prognostiziert werden, mit

[1] Zur besseren Lesbarkeit wurde in diesem Beitrag darauf verzichtet, eine Doppelung von weiblicher und männlicher Nennform anzugeben. Das jeweils verwendete Genus soll ausdrücklich beide Geschlechter einbeziehen.

welcher Wahrscheinlichkeit diese Person eine Aufgabe bekannter Schwierigkeit lösen wird, insofern Fähigkeit und Aufgabe den gleichen Kompetenzbereich betreffen. Dies wird als relatives Leistungspotential aufgefasst. Ein solcher Vergleich der Personen-Fähigkeit mit der Item-Schwierigkeit ist immer dann leicht möglich, wenn - wie z.B. in der Rasch-Modellierung - die Parameter θ_i und β_j auf der gleichen Achse abgetragen werden. Jedem Item ist hier eine charakteristische Item-Funktion zugeordnet, wobei die Schwierigkeit eines Items gerade dem Wendepunkt, d.h. der Stelle mit Lösungswahrscheinlichkeit von 0,5 entspricht.

Der statistisch-arbeitsmethodische Aufwand sämtlicher IRT-Modelle besteht nun in der Notwendigkeit, die beiden Parameter θ_i und β_j möglichst gleichzeitig aus den Einträgen der Datenmatrix schätzen zu müssen. Relevant hierfür sind einzig die ungewichteten Zeilen- und Spaltensummen r_i bzw. s_j. Das bedeutet v.a., dass die Wahrscheinlichkeit, mit der eine Person mit einer bestimmten Fähigkeit eine Aufgabe löst, nicht davon abhängt, welche Aufgaben konkret gelöst wurden (Eigenschaft der suffizienten Statistiken). Als Schätzverfahren werden beispielsweise die Maximum-Likelihood-Schätzung - in verschiedener Form - sowie Markov-Chain-Monte-Carlo-Methoden verwendet, wofür Statistiksoftware unerlässlich ist.

Die angegebene Modellgleichung (1) erfüllt schließlich die vier Annahmen der Rasch-Modellierung in der Form von Strobl (2010: 8f.):

[R1] Berücksichtigung der Personen-Fähigkeit

[R2] Berücksichtigung der Item-Schwierigkeit

[R3] Zunahme der Lösungswahrscheinlichkeit mit Zunahme der Personen-Fähigkeit

[R4] Sicherstellung der Intervallgrenzen einer Wahrscheinlichkeit.

Zentrale Voraussetzung des Rasch-Modells und der IRT ganz generell ist die *lokale stochastische Unabhängigkeit* aller Items. Es wird also mit möglichst hoher Präzission verlangt, dass die Lösung einer Aufgabe, die Beantwortung einer Frage oder allgemeiner die Erfüllung eines Items nicht von anderen Aufgaben, Fragen oder Items abhängig ist. Diese starke Voraussetzung ist bei der Testkonstruktion in der Regel zunächst nicht erfüllt und muss in Prätests überprüft werden. Methodisch werden ungeeignete Items daraufhin aussortiert oder ersetzt - es findet folglich eine statistische Bereinigung der expertis gesetzten Item-Vorauswahl statt, um soweit wie möglich die Voraussetzung zu garantieren. Eine Modellierung unter Zuhilfenahme der IRT ist somit nur dann möglich, wenn die Items eines Testinstruments prinzipiell angepasst bzw. weggelassen werden können (vgl. Rost 1999: 143). Auch bei der PISA-Studie wurden Prätests zur Überprüfung der Aufgaben in den einzelnen Kategorien eingesetzt, um ungeeignete (v.a. non-valid) Items auszuschließen. Dass dieser Arbeitsschritt im Zusammenhang mit der Item-Auswahl mit nennenswertem Aufwand verbunden ist, zeigt sich bei PISA an der mehrfachen Verwendung der als geeignet

bewerteten Aufgaben in den einzelnen Erhebungszyklen, die bis heute nicht vollständig veröffentlicht sind. Im Übrigen muss an dieser Stelle kritisch angemerkt werden, dass in den Technischen Berichten / technical reports der Studie unzureichend auf diesen Prätest eingegangen wurde (Adams & Wu 2002), woraus sich keine methodologischen Folgerungen für die Entwicklung von Erhebungsinstrumenten in der Kompetenzforschung ableiten ließen.

Das hier vorgestellte Verfahren zur Kompetenzmessung entgegnet der IRT insoweit, als dass es erstens auf den Wahrscheinlichkeits-Begriff verzichtet, sodass zweitens die Voraussetzung der lokalen stochastischen Unabhängigkeit formal fortfällt. Die oben genannten vier Annahmen [R1] - [R4] sollen unter einer geeigneten Anpassung erhalten werden genauso wie die Parameter θ_i und β_j, die sich ebenfalls aus den Einträgen der Datenmatrix als Randsummen r_i und s_j ergeben. Der hauptsächliche Unterschied zu allen gängigen IRT-Verfahren betrifft damit die linke Seite der Modellgleichung, an der gerade keine Lösungswahrscheinlichkeit, sondern einzig ein Zahlwert K_i stellvertretend für die im Rahmen des Kompetenztests gemessene Performanz steht. Obwohl damit auf den ersten Blick eine enge Beziehung zu IRT-Modellen aus dem bisher Gesagten hervorzugehen scheint, kann nicht deutlich genug herausgestellt werden, dass mit ihr tatsächlich kein probabilistisches Modell mehr vorliegt, dieses Messmodell also in der KTT unterkommt. Die Frage, warum bei der Entwicklung einer nicht-probabilistischen Modellierung derart auf die IRT abgehoben wird, mag berechtigt erscheinen. Als Erwiderungen können gelten, dass Kompetenzmessung außerhalb der in weiten Teilen der Praxis erprobten probabilistischen Testkonzepte inzwischen einer Begründungsbedürftigkeit unterliegt und weiterhin dieser Ansatz an den vier Annahmen [R1] - [R4] insoweit orientiert ist, dass einzelne Grundideen der IRT inkludiert bleiben.

6.1.2 Einführung der Kenngröße K_i

Wesentliches Argument für den Verzicht auf ein IRT-Modell ist das zur Messung diagnostischer Kompetenz verwendete Testformat, welches bei der Arbeit mit Fallvignetten aus drei Einzelkomponenten besteht: Format der Fallvignette, Formulierung der Aufgabenstellung und Antwortformat. Dass in diesem Kompetenztest ein schriftliches Transkript einer unterrichtlichen Situation in Form eines zusammenhängenden Textes analysiert werden sollte, gestattete den Probanden weitgehende Offenheit bei der Bearbeitung des Tests. Als festgelegt dürfen einzig die Fallvignette und die allgemein formulierte Aufgabenstellung angesehen werden. Bei der Auswahl der Testaufgaben muss in der Regel begründet werden, dass diese repräsentativ für das dem Kompetenzmodell zugrundeliegende Verständnis der latenten Variable sind. Die Schulleistungsstudie PISA hat bei ihrer Aufgabenauswahl für die einzelnen Kompetenzbereiche Lesen, Mathematik und Naturwissenschaft auf das Literacy-Konzept verwiesen (vgl. Kirsch et al. 2002: 14-17; Stanat et al. 2002: 6f.) und verdeutlicht damit, in welcher spezifischen Form diese drei Kompetenzen im Rahmen der Testung zu begreifen sind. Die Verwendung von Fallvignetten vereinfacht diese Not-

wendigkeit, da erstens im Rahmen eines Kompetenztests oft jeweils nur ein Fall zu analysieren ist und zweitens lediglich erläutert werden muss, in welchem Maße dieser berufsfeldtypisch ist (vgl. Dirks 2012a, Kap. II.4). Damit ist jedoch keineswegs ausgeschlossen, dass die jeweiligen Fallvignetten und die fallbasierten Diagnosen nicht auch bildungstheoretisch gerahmt werden (vgl. Dirks & Hansmann 2012, Kap. III.3).

Bislang wurden in der quantitativen empirischen Kompetenzforschung Aufgaben eingesetzt, die in abgestufte Teilaufgaben (Items) untergliedert waren, welche idealerweise nur ein eng begrenztes Antwortformat besaßen. Letzteres hat sogar positiven Einfluss auf das methodische Vorgehen bei der Absicherung der Güte des Testinstruments. Einem solchen fertigen Testheft geht realiter ein ausführlicher Prätest voraus, der v.a. anhand von statistischen Auswahlkriterien die Güte der vorgelegten Items prüfen soll, wobei ein geschlossenes Antwortformat nicht per se zu besseren Prätestbefunden führt, wohl aber deren Überprüfung erleichtert. Voraussetzung bei einer solchen Testkonstruktion ist die – inzwischen auch in Handbüchern zur Testtheorie nicht mehr erwähnte – Annahme, dass bei der Auswahl der Testitems aus einer unerschöpflichen Fülle heraus eine Entscheidung zugunsten der geeignetsten Items getroffen werden kann. Diese Annahme berührt nicht nur den Inhalt der Items sondern auch deren Formulierung. In aufgabenbasierten Tests entspricht also jede (Teil-)Aufgabe einem Item, das selbst bzw. deren Lösung wiederum unmittelbar Ausdruck der Personen-Fähigkeit ist. Betrachten wir in einem Leistungstest ein Item als Problemaufgabe, so ist es naheliegend die Schwierigkeit dieses Items anhand der Wahrscheinlichkeit, diese Aufgabe zu lösen, zu definieren und dann mittels der erhobenen Daten auch daraus zu schätzen (Spaltensummen).

Die Items eines auf der Analyse einer Fallvignette basierenden Tests sind jedoch nicht als eine zufällige, später statistisch optimierte Aufgabenauswahl zu begreifen, sie beziehen sich einzig auf den ausgewählten Fall. In diesem Projekt werden sie als fallbezogene Einzelinformationen in der Analyse der Fallvignette des Tests verstanden. Die Schwierigkeit eines Items muss dann auf die analytische Erfassung dieser Einzelinformation aus der Fallvignette heraus durch die Probanden bezogen sein, d.h. ein Item ist als erfüllt bzw. gelöst anzusehen, wenn die zugeordnete Information dem Fall entnommen wurde. Dem liegt damit wiederum eine Lösungs- bzw. Erfüllungshäufigkeit zugrunde, die sich ebenfalls aus der Spaltensumme ergibt. Zur Interpretation dieser Schwierigkeit wird der Begriff des *Grades der Informationsverknüpfung* B_j verwendet. Es lässt sich ansetzen: Je komplexer die Informationsstruktur in einer Fallvignette ist, je mehr Informationen diese enthält, je „tiefer“ eine einzelne Information im Text „verwoben“ ist und/ oder je größere Anstrengungen für das Inferiern nicht expliziter Informationen aus dem Kontext vermutet werden können, desto schwieriger ist es, diese Information aufzufinden und analytisch zu behandeln (s. Dirks 2012c, Kap. II.8). Aufgrund des aktuellen Forschungsstandes ist zu konzedieren, dass die Struktur von Informationen in einem Text

nicht pauschal erfasst werden kann und eine Maßzahl für die Komplexität nicht existiert.

Als Prädiktor für die latente Variable diagnostische Kompetenz wird ein Wert K_i verlangt, der Auskunft darüber gibt, in welchem Maße eine Testperson die Items mit verschiedenen Graden der Informationsverknüpfung, d.h. mit verschiedener Item-Schwierigkeit, aus der Fallvignette analytisch entnehmen kann. Abweichend von den Modellen der IRT werden keine unmittelbaren Zusammenhänge zwischen der Schwierigkeit einzelner Items und der ermittelten Fähigkeit des Probanden gefordert. Anstelle des Parameters θ_i tritt im neuen Modell der Parameter T_j auf, der als *Diversität der Personen-Fähigkeit* bezeichnet ist, sich aber wie in der probabilistischen Modellierung aus den Zeilensummen der Datenmatrix ergeben soll. T_j liegt analog zur IRT die Überlegung zugrunde, dass es für das Messmodell von Bedeutung ist, ob ein Item hauptsächlich von Personen, die insgesamte viele oder wenige Items berücksichtigen, erfüllt wird.

Zusammenfassend ergeben sich die folgenden Prämissen an K_i, für welche einzelne Annahmen des Rasch-Modells als Grundlage dienten:

[K1] Berücksichtigung des *Grades der Informationsverknüpfung* B_j

[K2] Berücksichtigung der *Diversität der Personen-Fähigkeit* T_j

[K3] Zunahme des Wertes K_i mit steigender Anzahl von erfüllten Items

[K4] Zunahme des Wertes K_i bei Erfüllung inhaltlich unabhängiger Items (bei konstanter Anzahl erfüllter Items).

Die letzten beiden Forderungen [K3] und [K4] sind Monotonieeigenschaften, wobei [K3] analog zur Annahme [R3] des Rasch-Modells verlangt wird. Diese Prämisse bedeutet jedoch nicht, dass der K_i-Wert einer Person mit beispielsweise a erfüllten Items über dem K_i-Wert einer anderen Person mit $a-1$ als erfüllt angesehenen Items liegen muss, sondern nur, dass der K_i-Wert mit jedem erfüllten Item streng monoton steigt. [K4] ergibt sich aus der komplexen Struktur von Informationen innerhalb der Fallvignetten, die im Folgenden noch ausführlicher behandelt wird.

6.1.3 Forderungen an den Item-Pool $\Im$ einer Fallvignette

Ausgangspunkt für die Konstruktion eines Item-Pools ist die Fallvignette, die in diesem Projekt einer schriftlichen Transkription einer unterrichtlichen Situation entspricht. Fallschilderungen beinhalten im Allgemeinen ein komplexes Netz an Teilinformationen, deren Zusammenfassung, gruppierende Bündelung und/oder Auslassung dem Leser einen scheinbar weiträumigen analytischen Umgang gestattet und ein einheitliches Auswertungsverfahren erschwert. Fragebögen und Testhefte umfassen im Gegensatz dazu eine wohlbedachte Auswahl von Fragen oder Aufgaben, deren Antworten und Lösungen vom Auswerter zu einem späteren Zeitpunkt überwiegend isoliert betrachtet werden

können. Weiterhin ist bei der Arbeit mit Fallvignetten der Umgang mit Texten in gesprochener oder geschriebener Form charakteristisch, deren Analyse in einzelnen Berufsfeldern wie der Lehrerbildung mündlich oder wie innerhalb dieses Kompetenztests schriftlich zu leisten ist, sodass das Antwortformat folglich einerseits unabgegrenzt (im Vergleich zu mehreren zu lösenden Aufgaben) und andererseits in Textform (anstatt im Layout eines Fragebogens) bleibt. Daher kann bei einem Item-Pool zu einer Fallvignette weniger von einer expertis erstellten Item-Auswahl gesprochen werden als vielmehr von einer Item-Abbildung aus dem jeweiligen Fall. Der Item-Pool ist bezogen auf die Arbeit mit Fallvignetten die Grundlage der nachfolgenden Auswertung und kann auch als neues Analogon zum in der Forschung bekannten inhaltsanalytischen Vorgehen (vgl. Mayring 2002) betrachtet werden.

Wird bei einer Kompetenzmessung mit einem offenen Antwortformat gearbeitet, dann müssen im Item-Pool alle erwartbaren Antworten als Items Berücksichtigung finden. Da die Aufgabenstellung gerade in der Analyse der Fallvignette besteht, sind alle Informationen aus der Fallschilderung als einzelne Items zu formulieren. Dieses Vorgehen wird notwendig, da das Erstellen von Diagnosen im Kompetenzstrukturmodell (KM) gerade so verstanden wird, dass alle(!) relevanten Bedingungsfaktoren zur Situation, zum Akteurshandeln und zu den entstandenen Effekten ermittelt und in Bezug zueinander gesetzt werden sollen (vgl. Dirks 2012b, Kap. II.5). Die *Vollständigkeit* erfordert, dass einerseits keine der Fragestellung genügenden Informationen aus der Fallvignette weggelassen werden, andererseits jedem Item ein Fallbezug zugewiesen werden soll, welcher erstens konkret am Text explizierbar ist (KM: Step I - III) oder zweitens in sachlogischen Bündelungen, Erweiterungen und Folgerungen bzw. in der Vervollständigung von gegebenen Informationen mit Hilfe von Brückenhypothesen und Plausibilitätsannahmen bestehen kann (KM: Step IV - V). Diese erwünschte Eigenschaft des Item-Pools wird *Vollständigkeit* genannt. Damit deutet sich bereits an, dass diese starke Voraussetzung einen hohen Arbeitsaufwand bei der Konstruktion eines Item-Pools abverlangt und gleichzeitig die Länge der Fallschilderung Einfluss auf die Handhabung des Testinstruments hat. Um diesem Problem zu begegnen, ist denkbar, die Aufgabenstellung auf Teile oder einzelne inhaltliche Aspekte der Fallvignette zu fokussieren. Die *vollständige* Erfassung des Falles im Item-Pool ist bei offenen Antwortformaten der entscheidende Vorteil der Testkonstruktion, der unter einer offenen Aufgabenstellung einzig das Nebengütekriterium der Testfairness sicherstellen kann; umgekehrt würde jeder Ausschluss eines Items zudem einer testtheoretisch plausiblen Bewertung der Testantwort widersprechen. Dabei muss insbesondere akzeptiert werden, dass eine Bereinigung des Item-Pools anhand statistischer Kriterien nicht(!) möglich ist. Das hat bedeutsame Konsequenzen für die etablierten Verfahren zur Bestimmung der Testgüte, der in dieser Modellierung auf andere Weise begegnet werden muss (s. dazu auch Dirks, Hansmann & Baumbach 2012, Kap. II.7).

Ein ideales Vorgehen bei Analyse einer Fallvignette beinhaltet das KM mit

seinem algorithmischen Diagnoseverfahren in fünf Steps, von dem auch eine plausible Struktur des Item-Pools ausgeht. Diese Eigenschaft ist durchaus kennzeichnend für Kompetenzmodelle (vgl. Abs 2007: 73). Der Einsatz eines offenen Antwortformats hat jedoch zur Folge, dass diese idealisierte Struktur den Testantworten nicht immanent ist, sondern diese vielmehr einer intuitiven, von den Probanden selbst gewählten Gliederung unterliegen. Die Lösung bzw. Erfüllung eines Items darf folglich nicht davon abhängig sein, an welcher Stelle es in der Testantwort schließlich vorliegt. Um eine Doppelung gleicher Textinformationen aus der Fallschilderung durch Zuweisung in zwei verschiedene strukturelle Bereiche des Item-Pools, hier z.B. dasselbe Item zu zwei verschiedenen Steps, zu vermeiden, wird die *Einzigartigkeit* jedes Items verlangt. Auf diese Weise wird die Objektivität der Auswertung insoweit verbessert, dass die Codierer lediglich entscheiden müssen, ob ein Item erfüllt wurde und nicht zu ergründen haben, an welcher Stelle der Analyse es eingebunden wurde.

Typischerweise werden die Fallvignette, kleinere Sequenzen aus ihr oder sogar Einzelinformationen von den Probanden in verschiedener Form gedeutet und bewertet, selbst wenn die gleichen Informationen aus dem Fall als Argumente eingesetzt werden. Vorstellbar ist weiterhin, dass ein Item in variierter Form in der Testantwort wiedergegeben werden kann. Um die notwendige Präzision der Item-Formulierung nicht zu erschweren, ist es für solche Fälle zweckmäßig, eine summarische Form von Items mittels einer *oder-Verknüpfung* zuzulassen. Diese soll immer dann für Items angewendet werden, wenn alternative Analyseansätze zu einer fallbezogenen Information vorliegen. Mit diesem Mechanismus können insbesondere unterschiedliche persönliche Wertungen bei der Testauswertung ausgeblendet werden, wenn diese nicht testrelevant sind. In der vorgestellten Erhebung zur diagnostischen Kompetenz von Lehramtsstudierenden fanden sich in den Analysen beispielsweise immer wieder Einschätzungen, in welchem Maße das Lehrerhandeln in der im Test eingesetzten Fallvignette Merkmalen „guten Unterrichts“genügt. Solche Äußerungen differierten über die gesamte Testpopulation und ließen im Prätest zunächst keine sinnvolle Berücksichtigung im Testergebnis zu. Unter Zuhilfenahme einer *oder-Verknüpfung* könnte beispielweise als Item formuliert werden, dass die Probanden das Lehrerhandeln bewerten, eine sehr positive Einschätzung unter Rückgriff auf den Aspekt $A1$ erhalten *oder* zu einer durchschnittlichen Beurteilung durch das Argument $A2$ gelangen. Alternative Analyseansätze liegen in diesem Beispiel vor, da sich beide Beurteilungen gegenseitig ausschließen (Komplementarität), sie somit nicht gemeinsam in einer Testantwort auffindbar wären, beide aber sachlogisch zulässig sind.

Diese Anforderungen an die Items in der Konstruktion eines Kompetenztests erscheinen dann vorteilhaft, wenn alle Items dichotom sind, der Auswerter also nur entscheiden muss, ob ein Item erfüllt ist oder nicht. Für die Datenmatrix bedeutet dies, dass für ein erfülltes Item in seiner Spalte in der zugehörigen Zeile des Probanden eine „1“eingetragen wird, ansonsten eine „0“. Prinzipiell ist eine Anpassung dieser Modellierung für graduelle Items auch möglich.

Für eine Strukturierung der Items im Item-Pool bietet sich die sozialtheoretische Fundierung des Kompetenzmodells an; in fünf verschiedenen aufeinander bezogenen Steps ist auf diese Weise ein algorithmisches diagnostisches Vorgehen zur Analyse der Fallvignette im KM dargelegt. Dieses *Cluster* lässt sich in der Kompetenzmessung nur bedingt einbeziehen, gerade weil dieses Verfahren in der Ersttestung von keinem der Probanden, in der Zweitestung dann von wenigen Probandengruppen beherrschaft, gleichzeitig aber nicht durchgängig angewendet wurde. Werden diesen Steps die unter dem Oberbegriff der soziologischen diagnostischen Kompetenz firmierenden Teilkompetenzen zugeordnet, dann zerfallen sie in zwei Gruppen: Step I - III und Step IV - V (vgl. Dirks 2012b, Kap. II.5), die fortan den beiden Subskalen entsprechen: Items der Steps I - III sind Skala 1, Steps IV - V Skala 2 zugewiesen.

Unbeantwortet ist damit zunächst noch die Frage, wie intuitiv von den Probanden Fallarbeit betrieben und ihre Analysen struktiert werden. Im Prätest konnte ermittelt werden, dass Probanden inhaltlich geschlossene Absätze bilden und innerhalb dieser Einzelabschnitte sowohl rein deskriptive, in der Struktur des KM also Items der Steps I - III, mit interpretierenden und inferierenden bis hin zu wertenden Passagen verbinden, wobei letztere dem Deutenden Verstehen (Step IV) und Verstehendem Erklären (Step V) innerhalb des KM zugehörig ist. Die Fallvignette wird also weder konsequent vom Beginn zum Ende, noch algorithmisch im Sinne des KM bearbeitet, sondern Teilbereiche, die sich offenbar aus der Spezifität des Berufsfeldes ergeben, nacheinander betrachtet. Als solche Teilbereiche finden sich z.B. eine Fokussierung auf das Lehrerhandeln, die methodische Gestaltung des Unterrichts, Disziplinstörungen seitens der Schülerinnen und Schüler, Lehrertypen etc. Gute Diagnosen in diesem Kompetenztest sollen demnach nicht einfach nur mit der Erfüllung möglichst vieler Items verbunden sein, stattdessen konkreter noch eine Vielzahl dieser inhaltlichen Bereiche erfassen und berücksichtigen. Aus diesem Grund wird jedem Item ein Platz in einem inhaltlichen *Cluster* zugeordnet (s. Dirks, Hansmann & Baumbach 2012, Kap. II.7) und dieses nach [K4] in die Messung integriert. Welche inhaltlichen Bereiche sich aus einer Fallschilderung ergeben, orientiert sich erstens am fallimmanenten *Kernproblem* und zweitens am Berufsfeld der Probanden; so ist durchaus in Betracht zu ziehen, dass bei derselben Fallvignette Probanden aus verschiedenen Berufsfeldern zu anderen Schwerpunktbildungen kommen. Mehr noch ist z.B. davon auszugehen, dass die methodische Gestaltung des Unterrichts insbesondere von Lehramtsstudierenden Gegenstand der Diagnose ist. Obwohl dieses inhaltliche Cluster aufgrund seines fallspezifischen Charakters im KM nicht verallgemeinert werden kann, ist es als Prämisse zum Maßstab in dieser Kompetenzmessung unerlässlich und kann gerade über diese Modellierung als Strukturmerkmal dieses Kompetenztests verankert werden. Insoweit trägt es zur Strukturierung des zugrundeliegenden Kompetenzbegriffes bei, da es konkret erklärt, wie kompetente fallanalytische Arbeit aussieht (vgl. Klieme & Leutner 2006; Schott & Ghanbari 2008: 22). Gemäß dem KM lassen sich die inhaltlichen Item-Gruppen als *Dimensionen des Kernproblems* einer Fallvignette verstehen und bezeichnen (vgl. Dirks 2012b, Kap. II.5).

Zusammengefasst lassen sich diese Grundsätze für die Konstruktion des Item-Pools einer Fallvignette als Voraussetzungen für die Anwendung des folgenden Modells formulieren:

[I1] Der Item-Pool einer Fallvignette ist *vollständig*.

[I2] Jedes Item ist *einzigartig*.

[I3] Komplementäre Items werden in *oder-Verknüpfungen* zu einem Item kontrahiert.

[I4] Jedes Item hat *dichotomen* Charakter.

[I5] Jedes Item ist nach seinem Inhalt in einem *Cluster* verortet.

Erwähnenswert ist ferner, dass im Item-Pool keine Items enthalten sein sollten, die auf fehlenden, nicht in der Fallschilderung enthaltenen Informationen beruhen. Diese dürfen als nicht erwartbar angesehen werden und sind daher redundant. Einzelne Ausnahmen sind mit Bezug auf das Berufsfeld der Probanden möglich, wie auch die Item-Pools zeigen (s. Dirks, Hansmann & Baumbach 2012, Kap. II.7). So konnten die Probanden aus den Fallvigetten z.B. keine konkreten Informationen zur Klassenstufe entnehmen, was von einer großen Mehrheit der Probanden festgestellt wurde, um davon ausgehend Inferenzen zu entwickeln.

Wenn oben ein am Inhalt der Items ausgerichtetes Cluster eingeführt wurde, welches gerade darauf beruht, dass Probanden intuitiv auf diese Weise ihre Analysen gliedern und bestimmte Items damit in eine argumentative Nähe rücken, dann wird plausibel, dass gerade keine Unkorreliertheit und stochastische Unabhängigkeit der Items zu verlangen ist. Auch wenn die Forschung für inhaltsanalytische Verfahren diese Voraussetzung gefordert hat (vgl. Holsti 1969: 95; Merten 1995: 98ff.; Atteslander 2008: 190), ist sie unter Beachtung aller Informationen aus der Fallvignette, d.h. unter der Annahme der Vollständigkeit, nicht sinnvoll, zugleich aber nicht mehr umsetzbar, da ein Ausschluss statistisch ungeeigneter Items nicht möglich ist. Als eine Art „Testgütekriterium“ist die *Vollständigkeit des Item-Pools* vorstellbar.

6.2 Messmodellierung

In Anknüpfung an die oben angeführten Modellannahmen soll nachfolgend ein konkretes Messverfahren mathematisch hergeleitet werden. Dabei ist zunächst die Kenngröße K_i in Abhängigkeit von den beiden Parametern B_j und T_j zu entwickeln. In der Praxis sind beide Parameter aus den beobachteten absoluten Häufigkeiten der erfüllten Items über alle Informanden bzw. alle Items zu ermitteln. Dass in die endgültige Berechnung des Kompetenzwertes K_i keine exakt übertragenen absoluten Häufigkeiten eingehen, sondern stattdessen auf kumuliert-geratete Werte verwiesen wird, ist mit den testpraktischen Schwierigkeiten der Auswertung geschlossener Textformate zu begründen. Auch unter genauer Formulierung der Items im Item-Pool muss davon ausgegangen

werden, dass die im Textfluss vorliegenden Testantworten der Probanden das Auffinden der Items für die Codierer maßgeblich erschwert und selbst bei intensiver Schulung einzelne Auswertungsfehler auftreten. Das unten dargestellte Rating-Verfahren, welches als zweite Komponente der Testdurchführung entstanden ist, kann als von den Modellannahmen unabhängig betrachtet werden. Dieser Algorithmus zielt v.a. auf die Verbesserung der Objektivität der Testauswertung, indem kalkulierbare Fehler bei der Codierung auf den Prädiktor K_i weniger Einfluss haben in Relation zur Verwendung exakter Beobachtungen. Der einhergehenden abnehmenden Genauigkeit der Messung ist Vorzug zu gewähren, da eine vermeintliche hohe Genauigkeit ohnehin durch eine geringe, aber nicht unbedeutende Zahl an unvermeidbaren Codierfehlern in Zweifel gezogen würde. Den Abschluss der Betrachtung bildet die Ausweisung von Kompetenzstufen, für welche die Spezifika der Analyse als Antwortformat aus vereinzelnten psycholinguistischen Erkenntnissen und den Befunden des Prätests zu verdeutlichen sind.

6.2.1 Herleitung der Modellgleichung

Zur Berechnung des Kompetenzstandes K_i einer Person i sollen die oben formulierten Modellannahmen [K1] bis [K4] verwendet werden. In [K1] und [K2] wird lediglich gefordert, dass K_i von den beiden Parametern T_j und B_j abhängig ist. Sowohl T_j als auch B_j sollen aus den gerateten absoluten Erfüllungshäufigkeiten des Items j durch Bildung von Reziproken bestimmt werden, also

$$T_j \in [0,1] \quad \wedge \quad B_j \in [0,1]. \tag{2}$$

Zur Einfachheit der Modellierung bewahren wir folgende Monotoniebeziehungen. Der *Grad der Informationsverknüpfung* eines Items j sei immer dann hoch, wenn dieses Item von wenigen Personen in der Fallvignette erkannt wird. Items, die oft als erfüllt vorliegen, zeichnen sich aufgrund der Reziprokbildung durch einen hohen B_j-Wert aus. Die *Diversität der Personen-Fähigkeit* nimmt dagegen zu, wenn v.a. Personen, die generell viele Items erfüllen, überwiegend auch dieses Item in ihrer Analyse berücksichtigen, sodass der T_j-Wert nahe bei 0 liegt. Mit diesen Begründungen und unter Nutzung der Monotonie der Exponentialfunktion wird der *Wert des Items j* definiert:

$$W_j := e^{B_j + [1 - T_j]}. \tag{3}$$

Diese Definition stellt die gewünschte Eigenschaft sicher, dass ein Item mit einem hohen Wert verbunden wird, wenn es erstens von Personen erfüllt wird, die insgesamt viele Items erfüllen (d.h. dieses Item eine hohe *Diversität der Personen-Fähigkeit* besitzt), und zweitens generell selten in den Analysen auffindbar ist (d.h. einen hohen *Grad der Informationsverknüpfung* aufweist). Wegen (2) und (3) ist dieses Produkt nach oben beschränkt und jedes Item j als Punkt $(B_j, 1 - T_j)$ in der Menge $[1,e] \times [1,e]$ darstellbar.

Nach [K3] soll weiter der Kompetenzstand einer Person i mit der Anzahl der

von ihr erfüllten Items zunehmen. Sei $u_{ij} \in U$ der Eintrag der Datenmatrix der j-ten Spalte und i-ten Zeile, dann kann unter Beachtung von [K1] bis [K3] mit $i = 1, \ldots, m$; $j = 1, \ldots, n$ gesetzt werden:

$$\overline{K_i} = \sum_{j=1}^{m} u_{ij} e^{B_j + [1 - T_j]} = \sum_{j=1}^{m} u_{ij} W_j. \tag{4}$$

Die Annahme [K4] kann nunmehr abschließend in die Berechnung von K_i einbezogen werden. Der gemäß [I1] bis [I5] erstellte Item-Pool wird mit Hilfe des Cluster-Verfahrens in disjunkte Item-Mengen P_l zerlegt, sodass eine Partition des Item-Pools $\mathfrak{I}$ als inhaltliche Untergliederung des Kernproblems dieser Fallvignette nach [I5] entsteht. Weiterhin gilt mit (2) bis (4):

$$1 + \overline{K_i} \geq 1 \quad \forall u_{ij} \in U. \tag{5}$$

Mit dieser Begründung liegt die endgültige Definition von K_i in der folgenden Form als Produkt nahe:

$$K_i := log \left[\prod_{l=1}^{w} \left[1 + \sum_{j \in P_l} u_{ij} e^{B_j + [1 - T_j]} \right] \right] \quad mit \quad P_1 \cup \ldots \cup P_w = \mathfrak{I}. \tag{6}$$

Anhand der Formulierung als Produkt von Summen zeigt sich der kompensatorische Charakter dieser Modellierung (vgl. Amelang & Schmidt-Atzert 2006: 399f.). Ein als konstant vorgestellter Prädiktorwert K_i kann durch ganz unterschiedliche Punktgewichtungen/-verteilungen in den einzelnen Faktoren, d.h. in der Fokussierung auf ganz verschiedene Dimensionen des fallspezifischen Kernproblems, gelingen. Für große Stichproben erscheint es sogar als zweckmäßig, diese Dimensionen mit Hilfe der Partition $P_1, \ldots, P_w$ aufzuschlüsseln. Dass tatsächlich keine lineare Kompensation zwischen $P_1, \ldots, P_w$ erfolgen kann, wird am Produkt-Operator sichtbar. Diese in [K4] geforderte Eigenschaft wird hierbei durch die *Ungleichung vom arithmetischen und geometrischen Mittel* (Königsberger 2003: 161; nach Cauchy 1821) garantiert:

$$\sqrt[k]{x_1 \ldots x_w} \leq \frac{x_1 + \ldots + x_w}{w} \quad \forall x_1 \ldots x_w > 0. \tag{7}$$

Mit äquivalenter Umformung der Wurzel auf die rechte Seite der Ungleichung (7) wird deutlich, dass ein Produkt mit konstanter Faktorenzahl w, deren Summe beschränkt festgelegt ist, dann besonders groß wird, wenn die einzelnen Faktoren ausgeglichen sind. Ein hoher K_i-Wert kann (unter festgehaltener Anzahl erfüllter Items) dann erreicht werden, wenn alle w Faktoren etwa den gleichen Wert besitzen, d.h. eine Person Items in verschiedenen Dimensionen des Kernproblems der Fallvignette in ihrer Analyse berücksichtigt hat. Eine Person mit der gleichen Anzahl erfüllter Items, welche jedoch überwiegend nur einige wenige dieser Dimensionen berühren, sodass mehrere Summen im Ausdruck (6) Null bleiben, wird einen niedrigeren K_i-Wert besitzen. Diese Eigenschaft lässt sich als *milde Kompensation* bezeichnen, da der kompensatorische

Einfluss auf den Prädiktor K_i gerade dann stärker zunimmt, wenn bereits weitgehender Ausgleich innerhalb der Partition $P_1, \ldots, P_w$ besteht, die Möglichkeit zu kompensieren also gering ist.

6.2.2 Bestimmung des Parameters B_j

Die zur Berechung des Kompetenzstandes K_i benötigten Parameter B_j und T_j sollen nun aus den Einträgen der Datenmatrix, v.a. den Spalten- und Zeilensummen s_j und r_i bestimmt werden. Zur Berücksichtigung der Schwierigkeiten bei der Auswertung von Analysen in geschlossenen Textformaten, die hauptsächlich ohne die bloße Suche nach Signalwörtern auskommen muss, soll ein Rating-Verfahren verwendet werden, das Auswertungsfehler in begrenztem Maße konzediert, indem es kleine Unterschiede der Beobachtungen tilgt. Zusätzlich kann es weiterhin in seiner Feinheit δ angepasst werden. Unter Verwendung der oben eingeführten Notationen und Variablen ergibt sich der Parameter B_j nach folgenden Rating-Algorithmus:

ST0 Seien s_j die Spaltensummen der Items $j = 1, \ldots, m$.
Bilde $\overline{S} = \{s_{(j)} \mid s_{(1)} < s_{(2)} < \ldots < s_{(\overline{m})}\}$.
Seien $\delta_s \in \mathrm{N}$ fest und $Z^s_{0,1} := [s_{(1)}, s_{(\overline{m})}]$.
Setze $k = 1$.

ST1 Bestimme $M^s_k := \{argmax\{|\, s_{(j)} - s_{(j-1)} \,| \geq \delta_s > 0 \mid s_{(j)} \in \overline{S}\} - \{s^1_{(j)}, \ldots, s^{k-1}_{(j)}\}\}$.

ST2 Falls $|\, M^s_k \,| = 0 \rightarrow STOPP$.
Falls $|\, M^s_k \,| = 1$, setze $s^k_{(j)} := s_{(j)} \in M^s_k \rightarrow$ ST5.
Falls $|\, M^s_k \,| > 1 \rightarrow$ ST3.

ST3 Bestimme $\overline{M^s_k} := \{s_{(j)} \in \{argmax\{Z^s_{k-1,v} \mid s_{(j)} \in M^s_k,\ v = 1, \ldots, k\}\}\}$.

ST4 Falls $|\, \overline{M^s_k} \,| = 1$, setze $s^k_{(j)} := s_{(j)} \in \overline{M^s_k} \rightarrow$ ST5.
Falls $|\, \overline{M^s_k} \,| > 1$, setze

$\overline{Z^s_{k-1,\bar{v}}} := argmin_{s_{(j)}}\{argmax\{Z^s_{k-1,v} \mid s_{(j)} \in \overline{M^s_k},\ v = 1, \ldots, k\}\}$

und wähle $s^k_{(j)} := s_{(j)} \in \overline{M^s_k}$ mit den Eigenschaften:
(1) $s^k_{(j)} \in \overline{Z^s_{k-1,\bar{v}}}$
(2) $s^k_{(j)} = argmin\{|\, s_{(j)} - M_{arith.}(\overline{Z^s_{k-1,\bar{v}}}) \,|\}$,
wobei $M_{arith.}(\overline{Z^s_{k-1,\bar{v}}})$ als arithmetisches Mittel aller s_j in $\overline{Z^s_{k-1,\bar{v}}}$ erklärt ist,
Falls aus (2) gilt: $s_{(j)} - M_{arith.}(\overline{Z^s_{k-1,\bar{v}}}) < 0$,
dann setze $t^s_k := s^k_{(j)} + \frac{1}{2}(s_{(j+1)} - s^k_{(j)}) \rightarrow$ ST6
Falls aus (2) gilt: $s_{(j)} - M_{arith.}(\overline{Z^s_{k-1,\bar{v}}}) > 0, \rightarrow$ ST5.

ST5 Setze $t^s_k := s_{(j-1)} + \frac{1}{2}(s^k_{(j)} - s_{(j-1)})$.

ST6 Bestimme die Zerlegung Z_k^s von $Z_{0,1}^s := [s_{(1)}, s_{(\overline{m})}]$:
$Z_k^s = [s_{(1)}, t_{(1)}^s] \cup \ldots \cup [t_{(k)}^s, s_{(\overline{m})}] = Z_{k,1}^s \cup \ldots \cup Z_{k,k+1}^s$.

ST7 Wenn $max\{Z_{k,w}^s \mid 1 < \mid Z_{k,w}^s \mid_{s_{(j)}}, w = 1, \ldots, k+1\} < k, \rightarrow STOPP$;
sonst setze $k := k + 1 \rightarrow$ ST1.

Sobald der Rating-Algorithmus gestoppt hat, soll für den Parameter B_j folgendes arithmetisches Mittel gesetzt werden:

$$B_j := \mid Z_{k,w}^s \mid \left[\sum_{s_j \in Z_{k,w}^s} s_j \right]^{-1} \quad \forall j = 1, \ldots, m; \;\; w = 1, \ldots, k+1. \tag{8}$$

Die Funktionsweise des Rating-Algorithmus' zielt zunächst nicht auf eine gleichmäßige Partionierung (gegeben durch die Zerlegung aus ST6) der Spaltensummen s_j, sondern auf eine Untergliederung nach dem Maßstab der Minimierung von Auswertungsfehlern. Aus diesem Grund wird in ST1 die Intervallabgrenzung stets an den Stellen des größten Abstandes der auftretenden s_j-Werte unternommen. Die STOPP-Regel in ST7 verhindert gleichzeitig zu große Intervalllängen.

Die besondere Gestaltung des ST4 entsteht aus der Betragsbildung in ST4 (2), an dessen Stelle explizit auf eine weitere Verkomplizierung der Formulierung verzichtet wurde. Daraus resultiert am Ende von ST4 eine Ausnahmeregel, welche wirksam wird, wenn eine negative Differenz im Betrag in ST4 (2) die neue Intervallgrenze t_k^s vom Mittelwert $M_{arith.}(\overline{Z_{k-1,\overline{v}}^s})$ künstlich entfernt, d.h. indem t_k^s in das links benachbarte Intervall $[s_{(j-1)}, s_{(j)}]$ verlegt würde, ohne dass das zuvor bestimmte arithmetische Mittel in diesem enthalten ist. Die Nähe der neuen Intervallgrenze t_k^s zum arithmetischen Mittel wird durch ggf. einsetzende Korrektur gesichert.

Für die Feinheit wird gewöhnlich $\delta_s = 1$ gesetzt. Wird dieser Wert erhöht, können auch gröbere Auswertungsfehler abgemildert werden, da mit ST2 ein rechtzeitiges Stoppen des Verfahrens erzwungen wird. Eine sinnvolle Festlegung der Feinheit δ_s ist stets abhängig von den auftretenden Werten s_j, welche wiederum in Relation zur Item-Anzahl m und zur Größe der Testpopulation n stark variieren können. Berücksichtigung sollte daher für alle die Abstände $a_j^s := [s_{(j)} - s_{(j-1)}]$, $j = 1, \ldots, \overline{m}$, die mit großer Häufigkeit $H(a_j^s)$ auftreten, die folgende Faustregel (9) finden:

$$\delta_s \leq a_j^s = [s_{(j)} - s_{(j-1)}] \quad \forall j = 1, \ldots, \overline{m} \;\; mit \;\; H(a_j^s) > \frac{\overline{m} - 1}{10}. \tag{9}$$

6.2.3 Bestimmung des Parameters T_j

Analog erfolgt die Bestimmung des zweiten Parameters T_j anhand eines Rating-Algorithmus' aus den Zeilensummen r_i. Hinsichtlich der Feinheit δ_z kann ebenfalls die Regel (9) verwendet werden.

ST0 Seien r_i die Zeilensummen der Personen $i = 1, \ldots, n$.
Bilde $\overline{R} = \{r_{(i)} \mid r_{(1)} < r_{(2)} < \ldots < r_{(\overline{n})}\}$.
Seien $\delta_z \in \mathrm{N}$ fest und $Z^z_{0,1} := [r_{(1)}, r_{(\overline{n})}]$.
Setze $k = 1$.

ST1 Bestimme $M^z_k := \{argmax\{\mid r_{(i)} - r_{(i-1)} \mid \geq \delta_z \mid r_{(i)} \in \overline{R}\} - \{r^1_{(i)}, \ldots, r^{k-1}_{(i)}\}\}$.

ST2 Falls $\mid M^z_k \mid = 0 \rightarrow STOPP$.
Falls $\mid M^z_k \mid = 1$, setze $r^k_{(i)} := r_{(i)} \in M^z_k \rightarrow$ ST5.
Falls $\mid M^z_k \mid > 1 \rightarrow$ ST3.

ST3 Bestimme $\overline{M^z_k} := \{r_{(i)} \in \{argmax\{Z^z_{k-1,v} \mid r_{(i)} \in M^z_k,\ v = 1, \ldots, k\}\}$.

ST4 Falls $\mid \overline{M^z_k} \mid = 1$, setze $r^k_{(i)} := r_{(i)} \in \overline{M^z_k} \rightarrow$ ST5.
Falls $\mid \overline{M^z_k} \mid > 1$, setze

$\overline{Z^z_{k-1,\overline{v}}} := argmin_{r_{(i)}}\{argmax\{Z^z_{k-1,v} \mid r_{(i)} \in \overline{M^z_k},\ v = 1, \ldots, k\}\}$

und wähle $r^k_{(i)} := r_{(i)} \in \overline{M^z_k}$ mit den Eigenschaften:
(1) $r^z_{(i)} \in \overline{Z^z_{k-1,\overline{v}}}$
(2) $r^z_{(i)} = argmin\{\mid r_{(i)} - M_{arith.}(\overline{Z^z_{k-1,\overline{v}}}) \mid\}$,
wobei $M_{arith.}(\overline{Z^z_{k-1,\overline{v}}})$ als arithmetisches Mittel aller r_i in $\overline{Z^z_{k-1,\overline{v}}}$ erklärt ist.
Falls aus (2) gilt: $r_{(i)} - M_{arith.}(\overline{Z^z_{k-1,\overline{v}}}) < 0$,
dann setze $t^z_k := r^k_{(i)} + \frac{1}{2}(r_{(i+1)} - r^k_{(i)}) \rightarrow$ ST6
Falls aus (2) gilt: $r_{(i)} - M_{arith.}(\overline{Z^z_{k-1,\overline{v}}}) > 0, \rightarrow$ ST5.

ST5 Setze $t^z_k := r_{(i-1)} + \frac{1}{2}(r^k_{(i)} - r_{(i-1)})$.

ST6 Bestimme die Zerlegung Z^z_k von $Z^z_{0,1} := [r_{(1)}, r_{(\overline{n})}]$:
$Z^z_k = [r_{(1)}, t^z_{(1)}] \cup \ldots \cup [t^z_{(k)}, r_{(\overline{n})}] = Z^z_{k,1} \cup \ldots \cup Z^z_{k,k+1}$.

ST7 Wenn $max\{Z^z_{k,w} \mid 1 < \mid Z^z_{k,w} \mid_{r_{(i)}}, w = 1, \ldots, k+1\} < k, \rightarrow STOPP$;
sonst setze $k := k + 1 \rightarrow$ ST1.

Sobald der Rating-Algorithmus gestoppt hat, soll für den Parameter T_j gesetzt werden:

$$T_j := \frac{1}{s_j} \sum_{i=1}^{n} u_{ij} \left[\mid Z^z_{k,w} \mid \left[\sum_{r_i \in Z^z_{k,w}} r_i \right]^{-1} \right] \quad \forall i = 1, \ldots, n;\ \ w = 1, \ldots, k+1. \tag{10}$$

Für den Sonderfall $s_j = 0 \ \ \forall s_j \in Z^z_{k,1}$, setze $T_j := 1$.

Der Vorteil der komplexen Formulierung der beiden Algorithmen zeigt sich darin, dass sie inhaltlich vollkommen übereinstimmen, sodass die Testauswertung lediglich die Kenntnis eines Rating-Verfahrens erforderlich macht. Deren

Anwendung im Rahmen der Messung diagnostischer Kompetenz von Lehramtsstudierenden gestattete die Auswertung verschiedenster Häufigkeitsverteilungen der Werte s_j (große Standardabweichung und niedrige Einzelhäufigkeiten) und r_i (niedrige Standardabweichung und hohe Einzelhäufigkeiten) mit demselben Rating-Verfahren.

6.2.4 Ermittlung der Kompetenzniveaus

Die Formulierung von Kompetenzmodellen umfasst in der Regel abschließend die Ausweisung von verschiedenen Kompetenzniveaus, welche den oft nur quantitativ vorliegenden Kompetenzstand (z.B. PISA-Punkte) wiederum in eine qualitativgestufte Beschreibung überführen, die in theoretisch fundierter Form kennzeichnen, welches Fähigkeitsspektrum eine Person mit dem entsprechenden Kompetenzstand tatsächlich besitzt. Wird die IRT zur Berechnung des Kompetenzstandes verwendet, dann liegt es nahe, die gemessene Fähigkeit einer Person anhand der gelösten Aufgaben und deren spezifischer Schwierigkeiten zu verifizeren. Bei der Arbeit mit Fallvignetten müsste jedoch analog dazu auch die Frage beantwortet werden, welches Maß die Schwierigkeit des Falles selbst angibt. Verschiedene Faktoren können die Schwierigkeit einer als Transkript vorliegenden Fallvignette determinieren:

- die Anzahl der Items
- die Komplexität der in der Fallvignette geschilderten Situation
- das kontextbezogene Vorwissen der Probanden
- das Vorwissen der Probanden hinsichtlich der verwendeten Textgattung bzw. des Testformats.

Unter Rückgriff auf die oben beschriebene Modellierung lassen sich insbesondere die beiden erstgenannten Faktoren in eine Maßgröße fassen. Das kontextbezogene Vorwissen der Probanden kann in der durchgeführten Erhebung zur diagnostisch soziologischen Kompetenz von Lehramtsstudierenden als vergleichbar angesehen werden. Zu berücksichtigen ist dabei die Fachsemesterzahl der Studierenden, wie in der Darstellung der Befunde eingehender thematisiert wird (s. Hansmann & Baumbach 2012, Kap. III.2).

Bei einer *vollständigen* Erfassung der in der Fallvignette geschilderten Situation im Item-Pool steht die Summe der Items insoweit stellvertretend für die Länge des Falltranskriptes, dass ausschließlich noch der Leseaufwand zum Beispiel durch Angabe der Textlänge zur sinnvollen Abschätzung der Bearbeitungszeit des Tests einzubeziehen ist. Dabei wird konzediert, dass das Verhältnis von Textlänge und Anzahl der Items im Allgemeinen nicht als konstant zu betrachten ist. So können schon kurze Fallvignetten zahlreiche Inferenzen erlauben, die sich als Items im Item-Pool abbilden. Dagegen ist genauso denkbar, dass knapp erfassbare Situationssequenzen durch weitschweifige Redewechsel gekennzeichnet sind, welche sich im Item-Pool zu wenigen Items verknappen.

Aus diesem Grund wurde in der vorgestellten Modellierung auf einen Einbezug der Transkriptlänge verzichtet und stattdessen verstärkt die Anzahl der Items als Einzelteile der Testantwort in den Blick genommen, um die Bearbeitungszeit für den Test festzulegen.

Wie lässt sich die Komplexität eines Falles in den Begriffen der vorliegenden Modellierung begreifen? Wird wieder von einer *vollständigen* Erfassung der Fallvignette ausgegangen, dann ist neben der Item-Anzahl m insbesondere der Wert jedes Items W_j eine relevante Kenngröße, da sie – aus den Zeilen- und Spaltensummen der Datenmatrix heraus – itembezogen Träger der Information ist, in welchem Maß Probanden mit verschiedener Personen-Fähigkeit dieses Item aus der Fallvignette analytisch entnehmen und in der Testantwort verarbeiten können. Zur Erinnerung: ein hoher Wert eines Items W_j steht für selten erkannte Items, die vorwiegend von Probanden mit insgesamt vielen Item-Treffern erfüllt worden sind. Die Komplexität des Falles drückt sich auch in Anzahl und Beschaffenheit der *Dimensionen des fallspezifischen Kernproblems* aus. Werden diese Cluster als inhaltlich verwandte Item-Gruppen aufgefasst, die für sich genommen mehrere algorithmische Steps im Sinne des KM berühren, dann erscheint jedes Cluster als eine eigene analytische Schleife, die von den Probanden im Test zu leisten ist. Intuitiv liegt es somit nahe, die Summe aller Werte W_j der Items unter dem Produkt der Clusterstruktur des Falles als Referenzgröße für die Schwierigkeit einer Fallvignette anzusetzen, was in der Modellierung mit (3) und $P_1 \cup \ldots \cup P_w = \mathfrak{I}$ dem maximalen K_i-Wert entspricht:

$$K_{imax} := log\left[\prod_{l=1}^{w}\left[1 + \sum_{j \in P_l} W_j\right]\right] = log\left[\prod_{l=1}^{w}\left[1 + \sum_{j \in P_l} e^{B_j + [1 - T_j]}\right]\right] \tag{11}$$

$$= log\left[\prod_{l=1}^{w}\left[1 + \sum_{j \in P_l} u_{ij} e^{B_j + [1 - T_j]}\right]\right] \quad mit \quad u_{ij} = 1 \tag{12}$$

Dieser Ansatz würde insoweit auch die bekannten psycholinguistischen Erkenntnisse zur Komplexität von Texten berücksichtigen (vgl. Dirks 2012a, Kap. II.4).

Für die abschließende Entwicklung von Kompetenzniveaus weisen die empirischen Befunde der Erhebung pädagogisch soziologischer Diagnosekompetenz von Lehramtsstudierenden insbesondere aus dem Prätest noch auf drei Besonderheiten hin: erstens zeigte sich in den Kontrollgruppen, dass die K_i-Werte von der ersten zur zweiten Testung deutlich anstiegen, obschon der zuerst eingesetzte Fall anhand Formel (11) nur etwa die Hälfte des K_{imax}-Wertes der zweiten Fallvignette besaß. Zweitens erreichten alle Gruppen stets einen gewissen K_i-Grundwert, der bei ernsthafter Fallanalyse offenbar auch ohne weitreichende diagnostische Fähigkeiten messbar ist. Und drittens finden sich einzelne Items, welche von keinem der Probanden des Prätests erfüllt wurden. Die ersten beiden Beobachtungen gehen vermutlich eng mit der deutlich variierenden Bearbeitungszeit in Einstiegs- und Abschlusstestung einher, die möglichst pro-

portional zu K_{imax} gehalten wurde. Unter testpraktischen Gesichtspunkten ist dieses Vorgehen plausibel, da ein Test, dessen Antwort mehr Teile in Form von einzelnen Items und Item-Clustern, somit einen größeren Umfang erfordert, unter Gewährung von entsprechend mehr Bearbeitungszeit stattfinden sollte. Ein solcher anteiliger Messansatz ist aus einer Modifikation von (11) zu erhalten, der als *Niveauwert des Probanden i* bezeichnet wird:

$$N_i := 100 \, \frac{K_i}{K_{imax}} \quad \forall i = 1, \ldots, n. \tag{13}$$

Die Plausibilität dieser Definition resultiert maßgeblich aus dem Einbezug der Bearbeitungszeit des Kompetenztests. Wird dagegen ein Vergleich zweier Fallvignetten mit unterschiedlichen K_{imax}-Werten bei gleicher(!) Bearbeitungszeit angestrebt, kann freilich diese Korrektur von K_i nach (12) in N_i bei der Ermittlung des Kompetenzniveaus des Probanden i entfallen.

Von Vorteil ist jedoch häufig die aus (12) folgende Eigenschaft der *fallunabhängigen Normiertheit*:

$$N_i \in [0, 100]. \tag{14}$$

Dieses Intervall kann nun gleichmäßig in verschiedene Abschnitte unterteilt werden, die den Kompetenzniveaus des KM entsprechen. Anhand der Formulierung des KM (vgl. Dirks 2012b, Kap. II.5) lässt sich eine Beschreibung der Kompetenzen nach den unterschiedenen Subskalen vornehmen (s. Dirks 2012c, Kap. II.8).

Werden wie in diesem Forschungsprojekt mehrere Subskalen unterschieden, dann ist mit (13) auf das Bewertungsmodell der *vektoriellen Skalen* (vgl. Hagemann 2002: 47ff.) zu verweisen. Die Messung pädagogisch soziologischer Diagnosekompetenz unterteilt nach dem KM zwei Subskalen: *Beschreiben von Situationsvariablen, Praktiken und Effekten* (Skala 1) sowie *Deutendes Verstehen und Verstehendes Erklären* (Skala 2). Sollen die beiden als orthogonal zu betrachtenden Subskalen in einem Maßstab zusammengeführt werden, dann entsteht ein quadratisches Niveaumodell. Dieses soll ausgeglichene Kompetenzwerte auf beiden Subskalen innerhalb seiner Niveaus begünstigen, d.h. eine bestmögliche Steigerung pädagogisch soziologischer Diagnosekompetenz soll entlang der Hauptdiagonale zwischen den definierten Subskalen stattfinden. Damit begründet sich auch der kompensatorische Charakter des Niveaumodells, dem durch senkrecht zur Hauptdiagonale verlaufende Niveaugrenzen Beschränkungen gesetzt sind. Dieser Ansatz ist mit der Normierung in (12) auf beliebig viele Subskalen erweiterbar. Auf eine Illustration wird an dieser Stelle verzichtet und auf die Passagen zur Testauswertung verwiesen (s. Hansmann & Baumbach 2012, Kap. III.2).

Literatur

Abs, Hermann Josef (2007). Überlegungen zur Modellierung diagnostischer Kompetenz bei Lehrerinnen und Lehrern. In: Lüders, Manfred &

Jochen Wissinger (Hrsg.), Forschung zur Lehrerbildung, Kompetenzentwicklung und Programmevaluation, Münster: Waxmann, 63-84.

Adams, Ray & Margaret Wu (2002). PISA 2000 technical report, Paris: OECD.

Amelang, Manfred & Lothar Schmidt-Atzert (2006)[4]: Psychologische Diagnostik und Intervention, Heidelberg: Springer Medizin Verl.

Atteslander, Peter (2008)[12]. Methoden der empirischen Sozialforschung, Berlin: Schmidt.

DeMars, Christine (2010). Item Response Theory. Understanding Statistics Measurement, New York: Oxford University Press.

Dirks, Una (2012a). Aufgabenformate – das Genre ‚Fallvignette'. In: Wilfried Hansmann; Una Dirks & Hendrik Baumbach (Hrsg.), Professionalisierung und Diagnosekompetenz – Kompetenzentwicklung und -förderung im Lehramtsstudium. (Kap. II.4, Online-Schriftenreihe der Philipps-Universität Marburg). [URL: `http://archiv.ub.uni-marburg.de/opus/schriftenreihen_ebene2.php?sr_i-d=30&la=de`].

Dirks, Una (2012b). Pädagogisch soziologische Diagnosekompetenz im Modell. In: Wilfried Hansmann; Una Dirks & Hendrik Baumbach (Hrsg.), Professionalisierung und Diagnosekompetenz – Kompetenzentwicklung und -förderung im Lehramtsstudium. (Kap. II.5, Online-Publikation der Philipps-Universität Marburg). [URL: `http://archiv.ub.uni-marburg.de/opus/schriftenreihen_ebene2.php?sr_id=30&la=de`].

Dirks, Una (2012c). Prä- und Posttest-Kongruenzen von Fallvignetten. Gemeinsamkeiten und Differenzen in der fallspezifischen Anforderungsstruktur. In: Wilfried Hansmann; Una Dirks & Hendrik Baumbach (Hrsg.), Professionalisierung und Diagnosekompetenz – Kompetenzentwicklung und -förderung im Lehramtsstudium. (Kap. II.8, Online-Publikation der Philipps-Universität Marburg). [URL: `http://archiv.ub.uni-marburg.de/opus/schriftenreihen_ebene2.php?sr_i-d=30&la=de`].

Dirks, Una & Wilfried Hansmann (2012). Die Ergebnisse aus bildungstheoretischer Perspektive. In: Wilfried Hansmann; Una Dirks & Hendrik Baumbach (Hrsg.), Professionalisierung und Diagnosekompetenz – Kompetenzentwicklung und -förderung im Lehramtsstudium. (Kap. III.3, Online-Publikation der Philipps-Universität Marburg).

Hagemann, Birte (2002). Leistungsmessung bei schriftlichen mathematischen Problemlösungen in Abhängigkeiten vom Bewertungsmodell, Diss. phil. Univ. Duisburg.

Hansmann, Wilfried & Hendrik Baumbach (2012). Entwicklung pädagogisch soziologischer Diagnosekompetenz von Lehramtsstudierenden.

In: Wilfried Hansmann; Una Dirks & Hendrik Baumbach (Hrsg.), Professionalisierung und Diagnosekompetenz – Kompetenzentwicklung und -förderung im Lehramtsstudium. (Kap. III.2, Online-Schriftenreihe der Philipps-Universität Marburg).

Hansmann, Wilfried; Una Dirks & Hendrik Baumbach (2012). Item-Pools der administrierten Fälle. In: Wilfried Hansmann; Una Dirks & Hendrik Baumbach (Hrsg.), Professionalisierung und Diagnosekompetenz – Kompetenzentwicklung und -förderung im Lehramtsstudium. (Kap. II.7, Online-Schriftenreihe der Philipps-Universität Marburg: Professionalisierung und Diagnosekompetenz).

Holsti, Ole R. (1969). Content analysis of the social science and humanities, Reading (Mass.): Addison-Wesley.

Kirsch, Irwin; John de Jong; Dominique Lafontaine; Joy McQueen; Juliette Mendelovits & Christian Monseur (2002). Lesen kann die Welt verändern. Leistung und Engagement im Ländervergleich. Ergebnisse von PISA 2000, Paris: OECD.

Klieme, Eckhard & Detlev Leutner (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen, überarbeitete Fassung des Antrags an die DFG auf Einrichtung eines DFG-Schwerpunktprogramms „Kompetenzdiagnostik“.

Klieme, Eckhard & Johannes Hartig (2008). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In: Manfred Prenzel, Ingrid Gogolin & Heinz-Hermann Krüger (Hrsg.), Kompetenzdiagnostik (= Zeitschrift für Erziehungswissenschaft, Sonderheft 8), 11-29.

Königsberger, Konrad (2003)[6]. Analysis 1, Berlin: Springer.

Mayring, Philipp (2010)[11]. Qualitative Inhaltsanalyse, Grundlagen und Techniken, Weinheim: Beltz.

Merten, Klaus (1995)[2]. Inhaltsanalyse, Einführung in Theorie, Methode und Praxis, Opladen: Westdt. Verl.

Pfeiffer, Dietmar K. & Carsten Püttmann (2006). Methoden empirischer Forschung in der Erziehungswissenschaft, Baltmannsweiler: Schneider-Verl.

Rasch, Georg (1960). Probabilistic models for some intelligence and attainment tests, Kopenhagen: Nielsen & Lydiche.

Rost, Jürgen (1999). Was ist aus dem Rasch-Modell geworden?, in: Psychologische Rundschau 50/3, 140-156.

Schott, Franz & Shahram Ghanbari (2008). Kompetenzdiagnostik, Kompetenzmodelle, kompetenzorientierter Unterricht: zur Theorie und

Praxis überprüfbarer Bildungsstandards, ComTrans – ein theoriegeleiteter Ansatz zum Kompetenztransfer als Diskussionsvorlage, Münster: Waxmann.

Stanat, Petra; Cordula Artelt; Jürgen Baumert; Eckhard Klieme; Michael Neubrand; Manfred Prenzel; Ulrich Schiefele; Wolfgang Schneider; Gundel Schümer; Klaus-Jürgen Tillmann & Volkmar Weiß (2002). PISA 2000. Die Studie im Überblick. Grundlagen, Methoden und Ergebnisse, Berlin: Max-Planck-Institut für Bildungsforschung.

Strobl, Carolin (2010). Das Rasch-Modell: eine verständliche Einführung für Studium und Praxis (= Sozialwissenschaftliche Forschungsmethoden 2), München: Hampp.

Zitation

Baumbach, Hendrik (2012): Das Messmodell. Messverfahren zur Analyse fallbasierter Diagnosekompetenzen [21 Seiten]. In: Wilfried Hansmann; Una Dirks & Hendrik Baumbach (Hrsg.), Professionalisierung und Diagnosekompetenz – Kompetenzentwicklung und -förderung im Lehramtsstudium. (Kap. II.6, Online-Schriftenreihe der Philipps-Universität Marburg). [URL: `http://archiv.ub.uni-marburg.de/opus/schriftenreihen_ebene2.php?sr_id=30&la=de`].