Social desirability in survey research: Can the list experiment provide the truth?

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

(Dr. rer. nat.)

dem

Fachbereich Psychologie

der Philipps-Universität Marburg

vorgelegt

von

Dipl.-Sozialwiss. Stefanie Gosen

aus Borken

Marburg/Lahn im Januar 2014

Diese Arbeit wurde gefördert durch ein Promotionsstipendium des DFG-Graduiertenkollegs „Gruppenbezogene Menschenfeindlichkeit“ (GRK 884) an den Universitäten Marburg und Bielefeld.

Social desirability in survey research: Can the list experiment provide the truth?

Dipl.-Sozialwiss. Stefanie Gosen

Am Fachbereich Psychologie der Philipps-Universität Marburg (Hochschulkennziffer 1080) als Dissertation am 27.01.2014 eingereicht.

Erstgutachter: Prof. Dr. Ulrich Wagner (Philipps-Universität Marburg)

Zweitgutachter: Prof. Dr. Peter Schmidt (Justus-Liebig-Universität Gießen)

Tag der mündlichen Prüfung (Disputation): 10.04.2014

Table of contents

List of tables

1. Introduction

"*Shying away from controversial topics, simply because they are controversial is an avoidance of responsibility." Sieber & Stanley (1988: 55)*

The number of studies that deal with controversial topics has fortunately grown in recent years. The research topics were further expanded for the use of such research surveys. This is not least due to the fact that by certain social events, new research fields were developed or rather received more attention (e.g., relation between terror attacks and anti-Islam attitudes, the increased fear of immigrants, psychological diseases, for instance, burnout). In the international context one can find surveys like the US National Survey on Drug Use and Health, which included questions about drug consumption or abortion, the US National Crime Victimization Survey (NCVS) with questions about criminal victimization, the US General Social Survey with questions about voting for a female president, or the Gallup Poll (Research Institute located in 27 countries), which included questions to race relations or opinions about foreign countries like Iran or Russia. Also in the German-speaking area, one can find surveys on extremely controversial and sensitive issues. For example, the General Social Survey (ALLBUS) contains questions regarding inequality, religion, deviant behavior, taxes and income; the Federal Center for Health Education (BZgA) included questions about AIDS prevention, prevention of sexual abuse and health promotion for people with migration background, the Politbarometer with questions about controversial political topics and attitudes, or also the interdisciplinary research project of Group-Focused Enmity (GFE; Heitmeyer, 2002), which surveyed attitudes in the area of discrimination, racism, anti-Semitism, etc. in an empirical long-term-study (see also Zick et al., 2008).

As mentioned before, all these surveys focus on controversial or 'sensitive' topics. Especially these sensitive topics lead to serious problems in surveys. Respondents tend to

bias their self-reports in a positive and admirable way, which leads to a systematic error when it comes to sensitive topics like racism. This systematic error is called social desirability bias and it is accountable for substantial distortions within surveys. That social desirability bias exists could be empirically proven in several studies, e.g., Edwards, 1957; Paulhus, 1984; Crowne & Marlowe, 1960; Aquilino, 1994; Smith, 1992; Tourangeau, Rips, & Rasinski, 2000. In order to counteract the problem of social desirability and to increase the response quality or rather to obtain more truthful answers from the respondents, researchers investigated and developed new possibilities. Next to social desirability scales, which were designed to measure social desirability (Paulhus, 1984), methods were developed in order to reduce the social desirability bias by increasing the anonymity while answering sensitive questions. One of these methods is the list experiment (Miller, 1984). The list experiment generates an estimated proportion of respondents who agree to one sensitive item on the aggregate level, which should increase the response quality. Only a few validity studies can be found, but there are several applications in which the list experiment proceeds as expected. There are some few studies in which the list experiment provides no results in the expected direction and failed completely to reduce the social desirability bias although the preparation was very thorough. Up to this point, it has been a matter of speculation why the list experiment failed in these studies, and it has not yet been empirically analyzed.

The present research deals with the validity of the list experiment as its main research question. In this respect, the effectiveness of the list experiment in prejudice research was analyzed. Furthermore, factors were found that might explain the inconsistent results of the list experiment.

In the following, the concept of sensitive questions will be described in more detail (Chapter 2). The following subchapter is dedicated to general survey errors, but especially to response errors that are often triggered by sensitive questions and concludes with the

theoretical conceptualization of social desirability response bias. Chapter 3 will give an overview of indirect measures and ends with a detailed description of the list experiment and the derivation of the research question. Furthermore, Chapter 4 deals with an introduction of the present research of this dissertation. This is followed by the two Manuscripts #1 and #2. Finally, this thesis is completed (Chapter 8 and 9) with a final discussion as well as an outlook to further research.

2. Sensitive questions in surveys

Before I will start to explain the concept of a 'sensitive question', I would like to give a short overview of the survey response process. Tourangeau, Rips and Rasinski (2000) divide the process into four components, and each of the components is linked to a specific mental process that contributes to answering a survey question. Table 1 describes the components plus the associated process. However, the authors do not expect that the respondents use all of the components and processes when they give a response in a survey. The model shows 13 different cognitive ways or possibilities to answer a survey question. Every decision that included the answer of a survey question depends on the personal and subjective factors of the individual, e.g., how accurate the answer should be, how long or short it takes to produce the answer and what the perceived consequences are. However, these factors that possibly contribute to generate an answer do not need lengthy deliberations (Tourangeau, Rips, & Rasinski, 2000: 14). The respondents produce an answer in less than 5 seconds (Tourangeau, Rips, & Rasinski, 2000).

Table 1 Components of the response process

Component	**Specific Process**
Comprehension	Attend to questions and instructions Represent logical form of question Identify question focus (information sought) Link key terms to relevant concepts
Retrieval	Generate retrieval strategy or cues Retrieve specific, generic memories Fill in missing details
Judgment	Assess completeness and relevance of memories Draw inferences based on accessibility Integrate material retrieved Make estimate based on partial retrieval

Response	Map judgment onto response category Edit response

Table is taken from Tourangeau, Rips, & Rasinski (2000: 8)

Nevertheless, all of the four specific processes can be interrupted or distorted by multiple factors, such as misinterpretation of a question, a threat to answer a question, forgetting crucial information, giving an answer to an inappropriate category, etc., and thus may lead to different response effects. A further important factor that influences the response process and belongs to the reporting errors is the motivated misreporting. This implies, for example, a social desirable answer to a social undesirable topic. This effect is one of the main points of this thesis. It emerges mainly when the topics within a survey are 'sensitive'. According to Tourangeau and Yan (2007: 876), within the response process, the motivated misreporting could appear in two different forms. On the one hand, misreporting could arise in the second category of "retrieval". During this process the respondents misreport by producing biased retrieval or by skipping this part completely. A social desirable answer might come about when respondents omit this part because they would not retrieve accurate information. Furthermore, it could also occur when respondents do not skip the "retrieval" part. In this process, the respondents retrieve certain information but distorted this information in such an extent that they present themselves in a positive way. On the other hand, misreporting could also appear in the last category (generate a response). In this case, the respondents produce a first answer on the basis of the previous processes. This answer or information, however, can be deliberately changed or edited in a socially desirable way by the respondents before uttering it (see Holtgraves, 2004 for a detailed description).

As mentioned in the introduction, questions about drug consumption, sexual behavior, discriminatory attitudes or voting behavior have something important in common – they are all sensitive topics (Lee, 1993; Tourangeau & Yan, 2007) and often cause distortions in

surveys. As mentioned above those topics cause difficulties within the 'normal' use of the response process. Even if the conceptualization of a 'sensitive question' seems obvious, finding a generalizable definition in the literature is rather difficult. The theory-driven concepts to explain the construct of 'sensitivity' range from investigations that refer solely to the researched topic up to definitions that comprehend the whole research activity and the implications for practice and the wider research community (Dickson-Swift, James, & Liamputtong, 2008).

Lee and Renzetti (1990: 512) provided a possible definition of a 'sensitive' topic, which, on the one hand, includes the aspect of threat and, on the other hand, they involve both the researcher and the researched. The authors defined a 'sensitive' topic as follows:

> a sensitive topic is one which potentially poses for those involved a substantial threat, the emergence of which renders problematic for the researcher and/or the researched the collection, holding , and/or dissemination of research data.

According to this definition, Lee (1993) pointed out that 'sensitive' research can be threatening in three ways. The first area is called 'intrusive threat' and implies areas like "private, stressful, and sacred" (Lee, 1993: 4). These areas often constitute an invasion of privacy, e.g., sexual or religious practice (Dickson-Swift, James, & Liamputtong, 2008; Wolf, 2012). The second area is the 'threat of sanction'. This kind of threat refers to studies of deviance and social control and it involves the possibility that investigations of specific issues may reveal stigmatizing and incriminating information (Lee, 1993: 4). An example of this constitute are Dickson-Swift, James and Liamputtong's (2008) interviews with drug addicts who may also show illegal behavior in form of drug-related crimes. The last form of threat, which often included controversies or involved social conflicts, is 'political threat'. In this field, it is difficult when it comes to political alignments "if 'political' is taken in its widest sense to refer to the vested interests of powerful persons or institutions, or the exercise

of coercion or domination" (Lee 1993: 4). For instance, political threat could occur if the majority of soldiers of the German armed forces agreed to right-wing extremist attitudes. This scenario might lead to problems of the image and authority of the German armed forces (Wolf, 2012: 29).

Important to note is that Lee and Renzetti (1990) took into account that other topics that are not 'sensitive' under normal circumstances can also be considered sensitive. Basically, this means that not exclusively the 'sensitive' character of the question is important, but also the influence of external factors or the interaction of question and social context in which the survey is conducted (Lee & Renzetti, 1990: 512).

Another approach to define the concept of 'sensitivity' from a social psychological perspective is presented by Tourangeau, Rips and Rasinski (2000; also Tourangeau & Yan, 2007). The authors involved the dimension of social desirability and distinguish between three distinct dimensions, two of which overlap with those suggested by Lee (1993). The first dimension is about 'intrusive' questions, which are perceived as too private or as taboo (Krumpal, 2013). These are questions that invade privacy and are out of bounds in everyday life, such as sexual behavior, personal finances, health status, etc. (Tourangeau, Rips, & Rasinski, 2000). Consequently, not only does the content of a question play a crucial role but also the situational environment or to whom the question is addressed (Tourangeau &Yan, 2007). In other words, Tourangeau and Yan described it as follows, "questions in this category risk offending all respondents, regardless of their status on the variable in question" (2007: 869). The second dimension includes the 'threat of disclosure'. The respondents are not able to answer a sensitive question truthfully because of the fear of possible risks, costs and consequences if information became known to a third person or institution (Krumpal, 2013). For instance, employees would never admit that they stole something from their own company because the consequences could be too serious (e.g., lose their job). Questions to

illegal behavior or incorrect behavior are considered to be potential factors that might lead to possible response errors in surveys. The last dimension and one of the principal foci of this dissertation concerning question's sensitivity is 'social desirability'. The concept of 'social desirability' implies that respondents adapt their responses to the social norm, which is specified by the society (Krumpal, 2013; Tourangeau & Yan, 2007). Respondents are motivated not to transgress social norms and to present themselves in a positive way. The reasons for this aforementioned motivation is grounded by the will of respondents to act conform to social standards, to create a positive self-image or to meet the expectations of an interviewer or a specific group. They want to receive the approval from the interviewer and at the same time they want avoid social sanctions and negative consequences that might be linked with a truthful answer (Tourangeau, Rips, & Rasinski, 2000). Consequently, the true attitude or the real behavior will not be an important part. The aim should be to adapt social desirable behaviors and attitudes and to refuse social undesirable behavior. The sensitivity of a question is often determined by answers of the respondents to specific survey questions. For example, a question about voting behavior is not sensitive, if the respondent votes (Tourangeau & Yan, 2007: 860). However, the question can be perceived as sensitive when the respondents do not vote, normally. Furthermore, social desirable responses can "also be conceptualized as respondents' temporary social strategies coping with the different situational factors in surveys (e.g., presence of interviewer, topic of questions, etc.)" (Krumpal, 2013: 2028).

In the literature there are two common approaches to 'social desirability' that try to explain the process behind (DeMaio, 1984). These two explanatory approaches show social desirability as a 'stable personality characteristic' and 'item characteristic'.

If social desirable response bias is characterized as a kind of 'stable personality characteristic', the focus will be on 'need for social approval' and 'impression management'

(Winkler, Kroh, & Spiess, 2006; Tourangeau & Yan, 2007; Krumpal, 2013). Thus, the respondent has a pronounced need for social approval and is determined to create a positive self-image. They provide no information that would shed negative light on them or respectively provide self-stigmatizing information. Crowne and Marlowe (1960) developed a measuring instrument for this construct in form of a scale based on behavioral items that are "culturally sanctioned and approved but which are improbable of occurrence." (Crowne & Marlowe, 1960: 350). In accordance to the critique that a one-factor scale does not cover the full construct, Paulhus (1984) generated a scale with a two-factor solution (Balanced Inventory of Desirable Responding). This scale is divided into 'impression management' and 'self-deception'. Impression management can be understood as a conscious and deliberate deception. The aim is to present oneself in positive light to the interviewer. Self-deception is used to protect self-esteem and self-image. It is a tendency to perceive the world in an optimistically distorted way. However, a certain degree of self-deception is typical even for a sane individual (Musch, Brockhaus, & Bröder, 2002).

The second dimension, which contributes to the explanation of social desirable response bias, is called 'item characteristic' (Phillips & Clancy, 1972). Basically, it means that the characteristic, the structure and the comprehensibility of the items, for example the topic that should be measure (i.e. the item content) as well as the presence of an interviewer, have an effect on social desirable responses. In other words, "[this] approach to social desirability response bias, perceived desirability of the item, considers behaviors or traits to be more or less socially desirable and thus discusses social desirability in relation to particular items" (Randall & Fernandes, 1991: 807). An example might be the investigation regarding abortion. It can be conducted to what extent this behavior is considered 'undesirable' by a single respondent or by the population or a specific group. However, it is

important whether the question captures the personal attitude of the respondent or of a specific group (e.g., Catholics).

In summary, it can be emphasized that both approaches to explain social desirable response bias should not be considered separately but together. The need for social approval, impression management and self-deception are aspects that are similarly responsible for social desirable response biases. Equally responsible are the perceived desirability of an item, the degree of privacy in an interview situation and the proportion of the population who behave socially undesirable (Krumpal, 2013). The following chapter will give a short overview of possible response errors and errors or rather biases that arise from 'social desirability'.

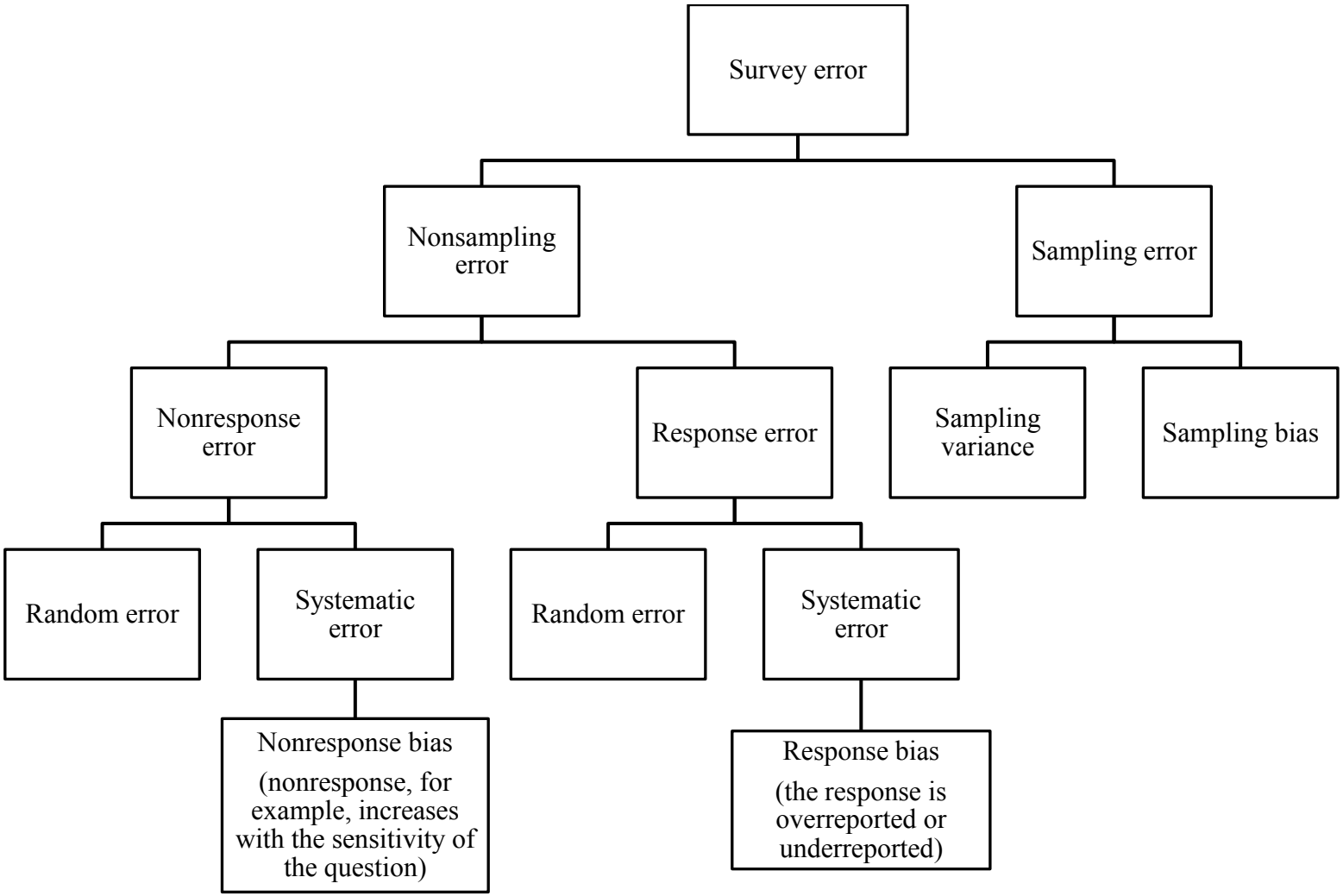
2.1 Sensitive questions and response errors

In general, surveys underlie different types of errors that substantially influence the data quality and the results of a survey, respectively. These errors occur among others when certain rules of the measuring accuracy are ignored or neglected. Hence, attention must be paid that the items have the same degree of difficulty (reliability). Items that are extremely difficult or extremely easy are less informative and should be not used, because they do not reveal differences between persons (distinction between persons with high characteristic values or low characteristic values). Moreover, items should have high validity to ensure that the construct measures what it is supposed to measure. Therefore, it is useful to take items that have a theoretical background and are empirically evidenced (Schnell, Hill, & Esser, 2005). In total, the literature shows two main survey errors that arise by conducting a survey (see Figure 1). These errors are known as sampling and nonsampling errors (Assael & Keon, 1982; Groves et al., 2004; Schnell, Hill, & Esser, 2005; see Krumpal, 2013 for an overview).

Usually researchers are conscious of the sampling error because it is not possible to measure all persons in a sampling frame[1]. Members of a target population are excluded deliberately from this group because of the selection. A sample is solely a selection of a representative group out of the target population that is used to measure its behavior, attitudes, characteristics, etc. in order to receive a picture of the whole society. The members of these groups have different values than the population. In addition, the sampling error can be split into two further errors. These errors are sampling bias and sampling variance. The sampling bias emerges when the sampling frame fails to represent certain members of the target population. A group might be underrepresented or not represented at all in the sampling frame because the possible set of selections excludes them systematically (Groves et al., 2004: 57). For example, a biased sample might be a survey of high school students that measures teenage use of Smartphone's but did not include students which receive education at home or dropouts like ill students. This means that the sample statistics deviate systematically from the statistics of the target population (see Groves et al., 2004 for a more detailed explanation).The sampling variance emerges when a sample frame is collected by randomization and thus contains many different subsamples of many different elements (e.g., districts, states, households, etc.). Each sample or subsample generates different values/estimates on the survey statistic. In order to measure the sampling variance, one should apply randomization and replication.

[1] A sampling frame is a set of persons of the target population that have a chance to be included into a survey sample.

Figure 1 Overview of the different survey errors.



Source: Own graphic

In surveys that seek out to collect data about sensitive issues, the nonsampling error is more common. This error can be divided into two subordinated errors, and each of these errors can be classified again into two parts. In the following, these errors and their negative influence on data quality will be described.

The first error is the nonresponse error. This error is not systematic and occurs when participants of a study are not present, refuse participation or do not fill out the questionnaire completely. The most important factor of a nonresponse error is "[that] values of statistics computed based only on respondent data differ from those based on the entire sample data." (Groves et al., 2004: 59). Furthermore and especially for sensitive topics or questions in surveys, the prevalence of the sensitive character is often underestimated, which leads to a detection rate on the lower limit of the truth prevalence (Ostapczuk, 2008). A reason for this problem might be the systematic nonresponse bias, which causes systematic distortions and thus systematic differences between nonrespondents and respondents (Groves et al., 2004; Krumpal, 2013). It is assumed that the nonresponse rate increases if the sensitivity of the question is increased. Tourangeau and Yan (2007) could show with the National Survey of Family Growth questionnaire that question sensitivity and nonresponse are positively associated (see Table 2). According to prior research (Juster & Smith, 1997), the item asking for total income presented the highest nonresponse rate. This is also due to the fact that this item can be seen in the broadest sense as very 'intrusive' (Tourangeau & Yan, 2007).

Table 2 Item nonresponse rates for the National Survey of Family Growth Cycle 6 Female Questionnaire, by Item

Item	**Mode of administration**	**% Nonresponse**
Total household income	ACASI	8.15
No. of lifetime male sexual partners	CAPI	3.05
Received public assistance	ACASI	2.22
No. of times had sex in past 4 weeks	CAPI	1.37
Age of first sexual intercourse	CAPI	0.87
Blood tested for HIV	CAPI	0.65
Age of first menstrual period	CAPI	0.39
Highest grade completed	CAPI	0.04

Note. ACASI = audio computer-assisted self-interviewing; CAPI = computer-assisted personal interviewing. The items are ranging from very sensitive to less sensitive. (Taken from Tourangeau and Yan, 2007: 862)

The other nonsampling error is called response error. This error is a random deviation and takes place when the true value from respondents differs from the measured values (Groves et al., 2004: 51–52). In case of the random response error, the values can vary unpredictably in repeated measurements. Therefore, the results (calculation of the mean level) of the independent repeated measures can be combined and the errors can cancel each other out. The response error, however, also includes a systematic error, which is called response bias. If respondents agree to answer a sensitive question; however, these responses are often not truthful or they are even euphemistical. The respondents deliberately misreport their answers to sensitive questions. Accordingly, the respondents corrected their answers up- or downwards. In other words, it basically means that respondents overreport socially desirable behavior and underreport socially undesirable behavior (Lee, 1993; Tourangeau & Yan, 2007; Krumpal, 2013). In general, respondents underreport topics like alcohol and

drug consumption (e.g., Sudman & Bradburn, 1974; Aquilino, 1994), criminal behavior (eg., Wyner, 1980) and unpopular attitudes, as anti-Semitism or racism (e.g., Dovidio & Fazio, 1992; Krysan & Couper, 2003; Krumpal, 2012). Furthermore, some studies could demonstrate that respondents overreport their responses when it comes to voting behavior (e.g., Holbrook & Krosnick, 2010a,b), seat belt use (e.g., Stulginskas, Verreault, & Pless, 1985), having a library card (e.g., Locander, Sudman, & Bradburn, 1976) or exercising (e.g., Tourangeau, Smith, & Rasinski, 1997). It should be noted that sensitive questions or nonexistent privacy are not only decisive to produce social desirability response biases. It also depends on different survey designs, strategies and the behavior of respondents' handling with sensitive questions. These possible behavioral patterns will be discussed briefly in the next chapter

2.2 Theoretical conceptualization of social desirability bias

This section will mainly discuss the rational side of misreporting or answering a sensitive question. The two main theoretical approaches are the rational choice theory (RCT) and the subjective expected utility theory (SEU). However, the theoretical basis or the theoretical precursors of the two main theories of social desirability are the social psychological concepts of symbolic interactionism (SI) and impression management which will describe briefly in the following.

2.2.1 Symbolic Interactionism (SI) and Impression Management Theory (IMT)

SI deals with the interaction between persons. It is based on the assumption that the meaning of social objects, situations or relations is a symbolic mediated process and that this process is created due to the interaction and communication with others. In general, three premises that were formulated by Blumer (1969) comprise the core of the theory.

1. "[…] human beings act toward things on the basis of the meaning that the things have for them, such things include everything that the human being may note in his world – physical objects, such as trees or chairs; other human beings, such as mother or a store clerk; categories of human beings, such as friend or enemies […]; guiding ideals, such as individual independence or honesty […]; and such situations as an individual encounters in his daily life."
2. "[…] the meaning of such things is derived from, or arises out of, the social interaction that one has with one's fellows."
3. "[…] these meanings are handled in, and modified through, an interpretative process used by the person in dealing with the things he encounters." (Blumer, 1969: 2)

According to Blumer, it can be summarized that human beings act because of meanings that they receive due to the interaction with others and the associated interpretation of the respective meanings.

In order to explain the action of a person in an interview situation and the possible resulting response bias, a specific model was developed, which is based on the concept of SI. In accordance with the scholars of SI and common action- or behavioral theories, Phillips (1971) stated that an interview for data collection is a special form of social action. In this situation, the respondents have the aim to control their impression in an optimal way. Thus, they want to receive a maximum of approval and personal satisfaction within the interview situation. In order to reach that goal, respondents try to analyze and interpret signals and expectations from the interviewer (communication partner) to weigh their alternatives. In this process, respondents give their answer according to the expectations of the interviewer and, on the other hand, according to their own aim. If there is a risk, for example, to get social rejection by a truthful answer, respondents would decide to take the easier way and adapt

their answer to the specific situation. According to Phillips, a truthful answer might only be possible if respondents feels that the subjective utility of a truthful answer will be higher than the consequences of an answer that they think is not expected by the interviewer and therefore might cause discomfort (Esser, 1985: 5).

The following part deals with Impression Management Theory (IMT). The theoretical basis of the IMT is derived from SI (Blumer, 1969) and other interaction theories. In general, impression management (IM; in social psychology also known as self-presentation; Schlenker, 1980; Tedeschi & Riess, 1981) assumes that human beings are acting actively and therefore also actively interact with their environment. This means that individuals influence their environment, their social surroundings, and their fellow human beings in an active and specific way. As in SI, the influence takes place by interaction processes. That implies the central assumption that human beings try to control or rather navigate consciously or unconsciously the impression that they give to others (Mummendey & Bolten, 1985: 57). Schlenker (1980: 6) pointed out that impression management is the "attempt to control images that are projected in real or imagined social interactions". In order to control and regulate the impression (self-presentation), individuals use different impression management strategies. In social psychology one often distinguishes between 'assertive' and 'defensive' techniques (short-term → specific situations) or strategies (long-term → across situations) (Jones & Pittman, 1982; Tedeschi, Lindskold, & Rosenfeld, 1985). Assertive techniques/strategies serve the individual by the use of interstratifications, flatter (ingratiate), proactive to receive advantages from other persons. Defensive techniques/strategies, however, rather serve as a defense in order to protect the identity of the individual (Mummendey, 2006: 52). One strategy, that would be included in the assertive strategies, is the 'ingratiation'. In this process the individual tries to ingratiate itself by flatter and laud (for example, compliments or opinion conformity). This occurs frequently when other persons

have a higher status as oneself (Schlenker, 1980; Aronson, Akert, & Wilson, 2004). A defensive strategy, which is well researched, is 'self-handicapping'. This strategy considers that the individual creates own obstacles to have an apology or an excuse for a personal failure (e.g., in a test) and accordingly cannot be held accountable for that (Kolditz & Arkin, 1982; Aronson, Akert, & Wilson, 2004: 179).

In the area of social desirability, impression management is a technique for respondents to receive social approval by giving the answer that is expected from the environment or from the interviewer to produce a positive self-image. Consequently, the respondents avoid negative reactions and sanctions by giving an adapted answer to their surroundings, also because they want to increase positive attention from the environment and especially from the interviewer (Krumpal, 2013). Respondents weigh their answers extensively to achieve these goals.

2.2.2 Rational Choice Theory (RCT)

In general the RCT anticipates that actors act on the basis of cost and benefit considerations. In other words, this is the choice between two available alternatives (Esser, 1975, 1986; Reinecke, 1991; Stocké, 2004). The RCT is one of the most popular theories in survey response behavior research, which takes into account and emphasizes the aspect of impression management in social desirability bias.

Especially in interview situations, the RCT comprises that respondents pursue the goal of choosing their answers to maximize social approval but simultaneously try to avoid repellent and sanctioned reactions from other people, such as the interviewer (Esser, 1975; Stocké, 2004; Stocké, 2007a; Krumpal, 2013). Respondents balance their truthful answer with the social desirable bias to create an expectation of how others will react to their answer, when they choose a response option (it is a combination of risks and losses) (Stocké, 2007a).

Stocké (2007b) presented three preconditions for the application of Rational Choice:

1. a strong desire for social approval,
2. a nonzero subjective probability of negative sanctions due to a perceived lack of privacy, and
3. respondents beliefs that the choice of one or another response option matters, i.e. that the other subjects' reactions will be clearly different for response option A compared to response option B. (Stocké, 2007b: 495 in Krumpal, 2013: 2031)

Only if all three preconditions are fulfilled, the multiplicative combination of the parameters reach enough impact to change the response behavior and consequently the related strength and direction of social desirability bias (Krumpal, 2013, Stocké, 2007b). Stocké (2007b) emphasized particularly that all three factors have to be fulfilled. If even only one of these conditions is not given, there is no effect to the 'prevalence of social desirability bias' and the respondents are willing to give a truthful answer. Stocké (2007b) could demonstrate in a study about racial attitudes that a three-way interaction between the aforementioned preconditions influences the responses to report their attitudes toward foreigners.

2.2.3 Subjective Expected Utility Theory (SEU)

An important variant of the RCT constitutes the SEU. The SEU (Savage, 1954) or respectively the behavioral model of the SEU is used to explain how individuals weigh their losses and gains to make risky decisions in interview situations. Based on this, it investigates in what way sensitive questions affect the decision of a respondent to give an honest or biased response (Tourangeau, Rips, & Rasinski, 2000).

Basically, it can be differentiated between two perceived factors, which contribute to a decision of a response: the rejection or agreement to a sensitive question. The first

perspective is called ‘perceived risks’. This perspective involves the perceived possibilities of alternative outcomes in consideration of each response option. The second factor, ‘perceived losses and gains’, combines the possible outcome with the respondent’s evaluation of this outcome (Rasinski et al., 1999; Krumpal, 2013: 2031). A truthful answer to a sensitive question can be compared to ‘perceived losses’. This might be embarrassment during the interview or disclosure of sensitive answers to persons or institutions. On the other hand, with a truthful answer, respondents are also able to generate ‘gains’. These ‘perceived gains’ might be approval from the interviewer, personal satisfaction or the promotion of knowledge about some topic or of public institutions (Tourangeau, Rips, & Rasinski, 2000; Krumpal, 2013).

Numerous studies have used the SEU to explain misreporting and biased responses to sensitive topics (e.g., Rasinski et al., 1994, 1999; Willis, Sirken, & Nathan, 1994). For example, Willis, Sirken and Nathan (1994) revealed in a number of studies that investigated the effect of social context and the data collection method concerning the motivation to answer truthfully in a survey, that the consideration of risks and losses regarding response disclosure yields a significant effect when it comes to the decision to answer a question truthfully (see Tourangeau, Rips, & Rasinski, 2000 for an overview of the studies).

In sum, the SEU is an interesting approach to measure and explain response behavior in general and misreporting in sensitive questions specifically. The respondents weigh their different losses and gains to calculate whether they give a truthful answer or not. This procedure is linked with various specific survey conditions (Rasinski et al., 1994; Krumpal, 2013). This means, for example that the researcher should provide a comfortable set for an interview, e.g., a high degree of privacy, in which the respondents are willing to give truthful answers to sensitive topics.

3. Methods to control and avoid social desirable response bias

In order to generate precise and 'truthful' information from respondents concerning sensitive topics, there are methods that increase the validity in self-reports and reduce the social desirability bias. The studies presented in this chapter do not imply the 'conventional' techniques (see Table 3 for a more detailed description), such as private setting (interviewer and bystander effect; Schuman & Converse, 1971; Aquilino, 1997), data collection mode (De Leeuw, 2001; Holbrook et al., 2003; Tourangeau & Yan, 2007) or question wording (Sudman & Bradburn, 1982; Fowler, 1995; Näher & Krumpal, 2012), but solely indirect and particularly questioning techniques.

Table 3 Description and examples of conventional question techniques

'Conventional' techniques	**Description**
Private setting	- respondents have reservations to reveal delicate information to the interviewer or third parties/bystanders (e.g., parents, siblings, friends, teachers, etc.) because they are afraid of negative consequences from these persons (Tourangeau & Yan, 2007) - in order to avoid this problem and to receive truthful information from the respondents, researchers try to create a private situation - for instance, the sealed envelope (e.g., Barton, 1958) technique is used in interview situations. During an interview, respondents receive a short questionnaire that has to be answered in a private setting without the interviewer. Afterwards the questionnaire is put in an envelope and therefore the interviewer has no idea about the answers.
Data collection mode	- in general there are three modes, which are interview-administered and are most utilized: paper and pencil personal interview (PAPI), computer-assisted personal interviews (CAPI) and computer assistant telephone interviews (CATI) - the most commonly used self-administered modes are: paper-pencil self-administered questionnaires (SAQ), computer-assisted self-administered questionnaire and web-surveys

	- for example, studies found that answers to sensitive questions are more honest/correct when the self-administered mode was used (see Tourangeau & Yan, 2007 for an overview of studies) - it was also found that in interview-administered questionnaires respondents tend to overreport socially desirable behavior (e.g., frequency of church going)
Question wording	- questions about sensitive topics should be formulated neutral, belittling and defusing to reduce the social desirable answer - the sensitive question should be unthreatening, euphemistic, familiar, and should included forgiving words and phrases (Krumpal, 2013: 2036) - one opportunity is the application of forgiving wording, apologizing or belittling questions. An example for such a question is: "Many Doctors now think that drinking wine reduces heart attacks and improves digestion. Have you drunk any wine in the last year?" (Sudman & Bradburn, 1982: 76) instead of "Did you drink wine regularly in the last year?"

Since the substance of this thesis is the validation and evaluation of an indirect survey method, the following subsections will give a short overview about different exceptional survey methods. It concludes with a detailed description of the list experiment, which is the central topic of this work, and the presentation of different studies.

3.1 The Bogus Pipeline Technique

The bogus pipeline technique was developed by Jones and Sigall (1971). This technique tries to reduce social desirable responses by having respondents believe that they are connected to an objective procedure (e.g., lie detector) that is able to show the interviewer the true score of their answer regardless whether respondents say the truth or not (Jones & Sigall, 1971; Tourangeau & Yan, 2007). In the proper sense, the respondents are consciously deceived into thinking that the researcher has an insight to their inner processes because of apparent electrodes, which are linked with the objective procedure.

If one connects the bogus pipeline technique with the SEU theory, the underlying principle of this technique is to increase the subjective costs of the respondents to misreport. In other words, it would be more embarrassing and unpleasant if the respondents appeared as liars than answering a delicate question truthfully (e.g., drug use, prejudice attitudes etc.) (Krumpal, 2013: 2037).

In the literature there is some empirical evidence that the bogus pipeline technique received significantly more reported socially undesirable attitudes or behaviors. A meta-analysis (Roese & Jamieson, 1993) about sensitive attitudes, such as racial prejudice, could indicate that the bogus pipeline technique showed significantly more socially undesirable attitudes. Other studies demonstrated that the bogus pipeline produced significantly more honest answers regarding sensitive behaviors like smoking, drug use, alcohol consumption etc. (Bauman & Dent, 1982; Murray et al., 1987).

In total, the bogus pipeline technique seems to be a promising method to reduce socially desirable responses. However, it is difficult to use this method in larger and national surveys because it is expensive, it takes time and it is rather complex.

3.2 Randomized Response Technique

The randomized response technique (RRT) introduced by Warner (1965) is an indirect survey method that produces and guarantees respondents privacy while they answer a sensitive question. The main principle of Warner's RRT is described as follows:

During the interview with the RRT, respondents receive two statements. One of the statements includes an undesirable topic (e.g., "I sometimes use marijuana."), and the other includes the negative form ("I never smoke marijuana."). Afterwards the respondent should give an answer to one question, which is selected by a randomizer (e.g., coin, dice, birthday), without revealing the answer to the specific question to the interviewer. The respondents should only answer with a "yes" or a "no".

On the basis of the given privacy, respondents are motivated to answer the sensitive question truthfully (Holbrook & Krosnick, 2010a; Lensvelt-Mulders et al., 2005). In general, it is possible to implement the RRT in many various forms (Coutts & Jann, 2011). Each of these different ways relies on a randomizing device (e.g., coins, dice, cards), which determines which one of the questions the respondent is asked to answer (Coutts & Jann, 2011; Krumpal, 2013). For instance, respondents could receive the sensitive question if they were born in May or June and get the negative form if they were born from July to April. Thus, the interviewer does not know which question is being answered, and therefore respondents are protected in their anonymity and privacy and will tend to give an honest answer. It is then possible, with a known random variable, to calculate an estimation of the prevalence of the socially undesirable behavior on the aggregate level. Aggregate level means that a value can only be calculated of the whole sample but not on the individual level as in the bogus pipeline technique or in social desirability scales. In the previously mentioned example, the estimation of the prevalence shows exactly how many of the respondents actually sometimes smoke marijuana (Lensvelt-Mulders et al., 2005Coutts & Jann, 2011;

Krumpal, 2013). In the literature there are numerous studies that corroborate that the RRT is working effectively and that provide more valid estimates regarding socially undesirable attitudes and behavior. Respondents who answered the RRT are more willing to report falsifying income tax reports (e.g., Himmelfarb & Lickteig, 1982), alcohol abuse (e.g., Volicer & Volicer, 1982), being prejudiced (e.g., Krumpal, 2012), etc. In addition, Lensvelt-Mulders et al. (2005) could demonstrate in a meta-analysis of 6 validation studies and 32 experimental studies (without validation studies) that the RRT received more accurate and truthful answers and reduced socially desirable response biases. Moreover, the authors found during the meta-analysis that dice and coins were mostly used as randomizing devices (Tourangeau &Yan, 2007). Next to the positive results of the studies, the RRT there are also some critical points. The RRT is not able to generate individual data. It produces only values or estimations on the aggregate level. However, with the development of logistic regression techniques it is possible to calculate correlations with background variables, but these correlations always involve a large error of the estimation (Lensvelt-Mulders et al., 2006). Another downside of the RRT is the intensive expenditure of costs and time. For instance, the rules of the RRT have to be explained to the respondents first. Lensvelt-Mulders et al. (2005) maintain that the RRT is more complex than direct questions and that the standard question response model (mentioned in Chapter 2; Tourangeau, Rips, & Rasinski, 2000) has to be extended. The four answering process factors of *understanding the question of being asked → retrieving the information from memory → integrating this information into summarized judgment → reporting this judgment correctly,* require a fifth factor for the RRT, namely respondents have to understand and follow the instruction of the RRT (Lensvelt-Mulders et al., 2005: 322). The cognitive burden increases with this kind of question format compared to standard question methods. An additional problem occurs when the respondents do not comply with the rules of the randomizer, thus do not act to the randomization principle.

Hence, the prevalence rate (estimation) of the socially undesirable behavior is underestimated and the advantages regarding direct self-reports decrease (Ostapczuk, 2008; Coutts & Jann, 2011).

3.3 The List Experiment

The list experiment is basically similar to the RRT. It also has the purpose of receiving a correct estimation of the sensitive topic by guaranteeing respondents privacy and anonymity. It was first presented by Miller (1984) as *unmatched count technique*, but it is also known as *item count technique* (Dalton, Wimbush, & Daily, 1994), *randomized list technique* (Zimmerman & Langer, 1995), *list randomization technique* (Karlan & Zinman, 2012), and *block total response* (Raghavarao & Federer, 1979). The procedure of the list experiment is quite easy and can be explained as follows:

The respondents of a survey are randomly split into two groups, which are called baseline and test condition. The baseline condition receives a list of items in which only nonsensitive items are included (this is the shorter list). The test condition receives the same list of nonsensitive items plus one sensitive item, the item of interest (this is, consequently, the longer list). Respondents in both conditions are requested to count just the number of items that they agree to. Thus, the interviewer is not able to find out which items the respondents chose and whether the sensitive item is included in their answer. Through this increase in anonymity, the probability of the respondents answering the questions truthfully increases. An example of the application of the list experiment from the field of racial prejudice, which nicely shows the calculations, is provided by Kuklinski, Cobb and Gilens (1997). In this study of the 1991 National Race and Politics Survey, the baseline condition was asked the following question:

> Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me how many of them upset you. I don't want to know which ones, just how many.
>
> 1. the federal government increasing the tax on gasoline;
> 2. professional athletes getting million-dollar salaries;
> 3. large corporations polluting the environment.

The test condition received the same question with nonsensitive items plus one sensitive item:

> Now I'm going to read you three things that sometimes make people angry or upset. After I read all three, just tell me how many of them upset you. I don't want to know which ones, just how many.
>
> 1. the federal government increasing the tax on gasoline;
> 2. professional athletes getting million-dollar salaries;
> 3. large corporations polluting the environment.
> 4. *a black family moving in next door.*[1]

In order to obtain an estimation of the proportion of people who were angry about the racial item – *a black family moving in next door* – we can calculate the mean level (ML) of reported items in both groups and then look at the difference between the two conditions. This calculation is only possible on the aggregate level. To obtain the difference, the baseline condition is subtracted from the test condition:

$$\hat{p} = ML_{TC} - ML_{BC}$$

TC = test condition (list with nonsensitive items plus a sensitive item)

BC = baseline condition (list with only nonsensitive items) (Tsuchiya, Hirai, & Ono, 2007)

To receive an estimate of the proportion of respondents, the mean difference has to be multiplied by 100 (Kuklinski, Cobb, & Gilens, 1997). In case of the racial prejudice study among southern residents, the mean level of the baseline condition amounted to 1.95 and the mean level within the test condition was 2.37. The difference between these two was 0.42 (the estimate of the proportion therefore was 42 percent). In other words, 42 percent were

[1] The sequence of the items is not fixed as it shown here. In most of the studies the list is randomized.

angered by the statement *a black family moving in next door* (Kuklinski, Cobb, & Gilens, 1997: 329).

However, in order to assess how the list experiment reduces the social desirability bias the estimation of list experiment is compared to direct self-report questions. In many studies is therefore used a "Difference of Proportion Test" that includes the *z-statistic:*

$$z = \frac{(\text{abs}(p_1 - p_2))}{\sqrt{(\text{SE}_1{}^2 - \text{SE}_2{}^2)}}$$

The standard error of the proportion of the direct self-report question is calculated as follow:

$$\text{SE}_\text{p} = \sqrt{\frac{(p * (1-p))}{n}}$$

Furthermore, the standard error for the list experiment is received from the difference between baseline and test condition (Holbrook & Krosnick, 2010b).[2]

3.3.1 Studies of the List Experiment

Later in this chapter, a variety of studies that demonstrate the effectiveness or ineffectiveness of the list experiment will be presented. There are, however, currently only a few studies that applied it in the field of prejudice research. In this chapter, I want to provide an overview and present three studies and their outcomes. Two of the studies deal with the election of a president in the US and the link attitudes towards a woman as president (sexism) and attitudes towards Jews (anti-Semitism). The third study is about the immigration restrictionism in the US. According to this research topic, it can be located in the area of xenophobia. The advantage of these three studies is that they used basically the same list experiment as Kuklinski, Cobb, & Gilens (1997). Kane, Craig and Wald (2004) conducted a

[2] There are also researcher that use multivariate regression techniques (e.g., Coutts & Jann, 2011; Blair & Imai, 2012; Glynn, 2013)

study about the attitudes or whether persons are angry about a Jewish candidate running for president (or vice president). The list experiment of Study 1 (vice president), showed a minimal difference between baseline and test condition of 3 percent (ns). Furthermore, in the second study (president), the difference amounted to 11 percent (ns). This value was higher than in Study 1 but again not significant. In this study, the authors received a nonsignificant result from the list experiment, which suggests that in this case it did not work correctly or rather as expected by the authors. The list experiment was not able to estimate the proportion of people who were angered by a Jewish candidate running for president. The second study was conducted by Streb et al. (2008) and it tested with use of the list experiment whether the respondents would vote for a female presidential candidate. Streb and colleagues compared the results within the list experiment to the findings of the national public opinion polls from the US to find out if these polls were influenced by a socially desirable response bias. The traditional polls indicated that only 5 to 15 percent would not vote for a female president. In contrast, the list experiment showed a significant difference between the test and baseline condition. It showed that about 26 percent of the respondents were angry about the idea of a female president. In accordance to the findings, the authors could demonstrate with the list experiment that the true attitude of the respondents who would not vote for a female president candidate was much higher than it had been assumed in traditional polls. The last one of the aforementioned studies by Janus (2010) regarding immigration restrictionism wanted to find out, as Streb et al. (2008), whether the general public polls were also affected by a socially desirable response bias. They could show with the list experiment[1] that the difference

[1] The list experiment was asked as follows:
Now I am going to read you three/four things that sometimes people oppose or are against. After I read all three/four, just tell me HOW MANY of them you oppose. I don't want to know which ones, just HOW MANY. Both groups are then given the same three nonsensitive items to choose from:
-The federal government increasing assistance to the poor.
-Professional athletes making millions of dollars per year.
-Large corporations polluting the environment.
-*Cutting off immigration to the United States.(sensitive item)*

between test condition and baseline condition amounted to 39 percent. That means that in total 61 percent (39 percent subtracted from 100 yields the unobtrusive estimate of support) of Americans agreed that immigration to the US should be cut off. In contrast, the direct self report question showed that only 42 percent of the respondents support immigration restrictionism. The study indicated a significant difference between list experiment and direct self-report question and suggests that many Americans did not say the truth about their attitude toward immigration restrictionism when they were asked directly.

In general, many other studies found that the list experiment produced higher estimates of the proportion of people for especially socially undesirable *behavior* than direct self-report questions. Dalton, Wimbush and Daily (1994) were able to show higher estimates within the list experiment in the field of unethical behavior (self-dealing) than in direct self-reports. Dalton, Daily and Wimbush (1997) found higher estimates of employee theft by using the list experiment compared to self-report questions. Also, LaBrie and Earleywine (2000) determined that the list experiment revealed higher estimates of sexual intercourse without condom and almost twice as high estimates of having sex without a condom after drinking, compared to direct self-reports. Another study provided by Tsuchiya, Hirai and Ono (2007) have shown that the list experiment had a higher estimate regarding shoplifting than the direct self-report question. There are several further empirical studies in which the list experiment yielded higher estimates than the direct self-report items. The following table gives an overview about studies in which the list experiment worked well.

Table 4 Overview of studies in which the list experiment received higher estimates than direct-self report questions

Study	**Question issue**	**Results**	
Dalton, Wimbush, & Daily, 1994	Unethical behavior	List experiment received higher estimates than direct self-report	
Wimbush & Dalton, 1997	Employee theft	List experiment received higher estimates than direct self-report	
Kuklinski, Cobb, & Gilens, 1997	Racism	List experiment received higher estimates than direct self-report	
LaBrie & Earleywine, 2000	Risky sexual behavior	List experiment received higher estimates than direct self-report	
Rayburn, Earleywine, & Davison, 2003a	Hate crime victimization	List experiment received higher estimates than direct self-report	
Rayburn, Earleywine, & Davison, 2003b	Anti-gay hate crime	List experiment received higher estimates than direct self-report	
Tsuchiya, Hirai, & Ono, 2007	Shoplifting	List experiment received higher estimates than direct self-report	
Streb et al., 2008	Sexism	List experiment received higher estimates than direct self-report	
Hoolbrook & Krosnick, 2010b	Voting behavior	List experiment received higher estimates than direct self-report	Online Survey: Direct self-report question received higher estimates than list experiment

Janus, 2010	Immigration restrictionism	List experiment received higher estimates than direct self-report	
Coutts & Jann, 2011	Drug consumption & infidelity	List experiment received higher estimates than direct self-report	

Source: Own table.

Next to these studies, a meta-analysis of the list experiment was conducted by Tourangeau and Yan (2007). They compared seven studies, only one of which was a general population survey that yielded very negative results. The direct self-report questions received higher estimates than the list experiment (Droitcour et al., 1991). The further studies were all undergraduates or other subsamples, like auctioneers. In sum, across all studies the authors found a small positive effect that the list experiment received higher estimates, but it was not significant. Unfortunately, this analysis has a very small sample size, which makes it very difficult to formulate a statement about the effectiveness and functionality of the list experiment in the literature. In this meta-analysis and in the previously described studies, the list experiment yielded useful results. The next part is devoted to special studies in which the list experiment was not able to provide valid information. These studies should give an insight in the arbitrariness of the results and a possible publication bias.

As mentioned above, Droitcour et al. (1991) conducted a general study to intravenous drug consumption and passive anal intercourse in the National Household Seroprevalence Survey Pretest (N = 1435). In this study, the list experiment was not able to generate a higher estimate than the direct-self reports. The authors can only presume why the list experiment did not produce efficient results in the study of drug use.

On the one hand, they suspect that the method might be difficult to answer. For example, the respondents could have made errors summing up the number of behaviors they have engaged in, e.g., they could have accidentally given the number of one particular item instead of the number of items they agreed to. If an item is placed on position three, the respondents might have indicated the position three and not the number of items they agree to. A further source of error might be that respondents have difficulties with cognitive processes associated with deciding which items they agree to. They also might have had trouble keeping track of the number of items while in the process of decision making. This might cause problems like forgetting an item or taking an item twice.

On the other hand, the researchers assumed that there are some discrepancies between the nonsensitive items and the sensitive item. The nonsensitive items are supposed to be "neutral" and not too conspicuous (low prevalence), but they also have the task to fit in the context of the sensitive behavior in question. Thus, it is very difficult to find the right nonsensitive items. Moreover, the respondents might become insecure answering the list experiment if they realize to big of a contrast between the nonsensitive items and the sensitive item. Biemer et al. (2005: 150) state that "[it] could make the respondent suspicious that there was some trick involved and that the investigators would in some way be able to determine whether the respondent had engaged in the sensitive behavior. Consequently, some respondents may have deliberately misreported the number of behaviors that applied."

Another study by Ahart and Sackett (2004) investigating counterproductive behavior applied five different sensitive items, like "I have stolen more than 5$ from a past employer", "In the past, I have called in sick when I wasn't actually ill", "I've done slow, sloppy work on purpose". They set out to find out if the list experiment produced higher estimates than the direct self-reports. The results (participants were psychology students, N = 550) indicated that the list experiment could not receive a significantly higher estimate than direct-self

reports in any condition. The authors provided that a possible explanation of the inconsistent results of this study could be the effect of the sample size. They assume that with a larger sample size the list experiment could have obtained more valid estimates. A further limitation of this study is that the respondents were undergraduates and not comparable to employees.

Finally, Biemer et al. (2005) (Biemer & Brown, 2005) also could not obtain any valid results from their application of the list experiment. It was conducted as part of the National Survey on Drug Use and Health (N = 70.000) to estimate the proportion of respondents' cocaine consumption in the past year. Before the actual survey could start, the authors did many further steps in order to guarantee the success of the list experiment and of the study, respectively. First, they conducted cognitive laboratory experiments to find out the optimal length of the list (ideal number of items in the baseline condition is four). Second, they did some more cognitive research to determine the best content of nonsensitive items. Their goal was to create items that strike a balance between being too nonsensitive and too sensitive and therefore threatening and prone to socially desirable responses. This seemed especially important in the context of drug abuse research. After the cognitive experiments, the final introduction and item list was as follows:

How many of the things on this list did you do during the past 12 months, since (date fill)?

Rode with a drunk driver;
Walked alone after dark through a dangerous neighborhood;
Rode a bicycle without a helmet;
Went swimming or played outdoor sports during a lightning storm;
Used cocaine, in any form one or more times.

The answer possibilities were:

0 = none of these things, 1 = one of these things, 2 = two of these things, 3 = three of these things, 4 = four of these things, 5 = five of these things.

Furthermore, they specified the sample size with simulation studies and found out that 35,000 responses was the minimum sample size to reach the precise level of accurate estimates. In addition, they also included the nonsensitive items (the short list) of the list experiment as direct questions in the main questionnaire to test the response consistency of the short list of the list experiment items and to evaluate the measurement error within the list experiment questions. The authors called this procedure "pseudo-IC" (item count technique) in which the indication of the number of *yes* answers could be produced for each respondent. On this basis and by using test-retest reliability, it might be possible to test the reliability of the list experiment questions.

Finally, the results of the analyses were not promising, despite of the meticulously preparatory work that had been done. In sum, the direct-self report questions yielded higher estimates of cocaine use in the past year than the list experiment. In this case, the list experiment failed completely to show higher estimates of cocaine use because it partly produced negative estimates. Why this study failed to such an extent or rather the list experiment did not show any effect can only be speculated by the authors, once again. One reason might be that the estimator of the list experiment is more biased than the direct-self report questions. In other words, the list experiment might have lead to specific measurement errors due to its complexity and the respondents' concerns about privacy (Biemer & Brown, 2005: 306). One solution to this problem, which the authors recommend, is to also ask the items of the list experiment within the baseline and test condition directly. These can be taken into account to correct the measurement error of the list experiment. Another explanation of the failure might be that the question of the list experiment was answered very unreliably. This was checked by comparing the list experiment with direct questions of the short list (mentioned above as "pseudo IC"). The reason might be that "the respondents' failing to give

careful thought to each item in the IC list when counting the number of applicable behaviors." (Biemer et al., 2005: 173)

In total, the literature regarding the list experiment shows inconsistent results. Many studies presented positive results, which at first sight support a valid application of the list experiment. On the basis of these list experiments that worked well, the authors established some guidelines to which one should pay attention when developing a list experiment. With the assistance of these criteria, they state it could be expected that list experiments will proceed appropriately and effectively.

In the following the conditions/factors under which the list experiment should work are listed:

- Nonsensitive items should have an adequate item difficulty. They should not generate too much agreement or too much rejection in order to avoid ceiling and floor effects;
- The nonsensitive items should be clear and cause strong opinions;
- The list of nonsensitive items should not be too short or too long. If the list is too short, respondents tend to underreport their answers because their anonymity is not longer fulfilled. If the list is too long, the nonsensitive items generate additional, irrelevant variance and deteriorate the effectiveness of the estimate of the proportion of the sensitive item. Therefore, the reason is that the sensitive item is only requested in one sample, which produces more variance because of the higher number of items (see Glynn, 2013). Moreover, measurement errors can occur because respondents are not able to remember their answers to all items. → Recommended are four items in the short and five items in the longer list;
- Direct self-report questions of the sensitive item should be included into the questionnaire to detect social desirable bias. (Droitcour et al., 1991; Tsuchiya, Hirai, & Ono, 2007; Blair & Imai, 2012; Glynn, 2013)

Analogous to this list, which contains criteria for the successful application of the list experiment, authors of studies that could not provide valid results tried to find reasons why the list experiment failed to produce higher estimates for the sensitive item. These are, however, speculations that do not constitute solid evidence. These conditions can be listed as follows:

- An error source could be that the respondents might have difficulties with the cognitive process of deciding which items they agree to and further difficulties in summing up or computing the number of items they agree to (Droitcour et al., 1991);
- There could be some discrepancies between the nonsensitive items and the sensitive item. Respondents might feel uncomfortable answering the list experiment when they realize to big of a contrast between the nonsensitive items and the sensitive item (Droitcour et al., 1991);
- The estimator of the list experiment might also be more biased than the direct-self report questions. In other words, the list experiment could be a procedure that leads to specific measurement errors due to its complexity and the respondents' concerns about privacy (Biemer et al., 2005);
- The questions of the list experiment could have been answered very unreliably. (Biemer et al., 2005)

As can be seen from these two lists and from the studies that were described beforehand, the list experiment suffers from inconsistent results, which question its validity. Even precise preparation and accurate development could not resolve these problems.

This leads me to one last point that further adds to the already mentioned critique of the list experiment – the publication bias. There is the possibility that studies with positive results/effects are published more often than studies with non-effective significant results (Rosenthal, 1979; Fanelli, 2012). Consequently, there might be many more unpublished

studies that obtained inconsistent results, and we might underestimate the amount.[2] It could very well be that the effect of inconsistency would actually be higher if one included the non-published studies and the error rate of significance tests.

Unfortunately it is therefore rather difficult to find studies that yielded negative results and view the list experiment critically. Most of the studies that used the list experiment – even the ones that failed – evaluate the technique positively and advise to conduct further studies; many view it as a promising technique that needs to get more attention in the research of sensitive topics.

In sum, it is not easy to find critical points and it is also difficult to find explanations for the failure of the list experiment that are supported by empirical evidence. The problem that arises for the research is that researchers that use it as a means of finding out more about sensitive topics might find it apparently promising, but it has – very likely – only survived due to random results and a publication bias that favors positive outcomes.

[2] By now there were developed special (peer-reviewed) Journals which deal with negative research results, like *The All results Journal, Journal of Articles in Support of the Null Hypothesis, Journal of Negative Results in Biomedicine*, *Journal of Contracting Design in Science*. In addition, one found also internet based archives that published and discussed for example, replication attempts in experimental psychology studies (http://www.psychfiledrawer.org/).

4. The present research

Based on the aforementioned findings and problems of the list experiment and also because of the lack of validation studies, the main goal of this work is to evaluate, validate and to find specific marginal conditions that might explain the inconsistent results of this technique. Furthermore, the efficiency of the list experiment should be diagnosed in terms of socially desirable responses on prejudice attitude items. Manuscript #1 studied the question of the validity of the list experiment and started out to find conditions that might explain the inconsistent results of the list experiment. The second Manuscript (#2) also tested the validity of the list experiment, but it utilized qualitative and experimental designs to find factors that provide details about the reasons for the inconsistent results.

Manuscript #1 dealt with the question of the validity and possible conditions of the list experiment's inconsistency. Since the list experiment has not been conducted sufficiently frequently in the field of prejudice (Kuklinski, Cobb, & Gilens, 1997; Kuklinski et al., 1997; Kane, Craig, & Wald, 2004; Streb et al., 2008), in this manuscript the list experiment was studied with two prejudice attitude items; these items came from the fields of anti-Semitism and Islamophobia. First, an adequate data basis was needed in order to test the external validity.

Manuscript #1 conducted and compared three different studies implementing a total of five list experiments, two different types of survey modes and a panel analysis. The first two studies were based on a representative sample (CATI) and included one list experiment each. Both studies contained the sensitive item regarding anti-Semitism. This item was also conducted as direct self-report question to detect social desirability biases in comparison. In this manner, approval rates could be compared between the baseline and test condition of the list experiment, and they could also be compared to the direct self-report questions. A further

online study was conducted to test the external validity and to have the possibility to compare several list experiments. This study included three different list experiments and also the suitable direct self-report items to allow a comparison of social desirability response bias. Here, the sensitive prejudice attitude items were about anti-Semitism and Islamophobia. Another list experiment that comprised only nonsensitive items was implemented to investigate whether the change or increase in the mean was caused by the higher number of items in the test condition, which was a further central question of the study. The results were then compared to the baseline condition. In order to ensure the accuracy of the results from the previous studies and to test the temporal stability of the list experiment, Study 3 used a panel design.

While the main focus of Manuscript #1 was on the external validity of the list experiment in which the factors that might explain the inconsistent results were only analyzed marginally, Manuscript #2 concentrated on the cognitive processes and two more factors that are responsible for the inconsistent results and the reason of the failure of the list experiment. One of the main research questions dealt with the shifted response patterns of the respondents due to the sensitive item in the test condition. In other words, when a sensitive item was included, the respondents moved their responses to the nonsensitive items because these were easier to answer. Study 1 conducted cognitive interviews to understand the cognitive processes of the respondents, how they answered a list experiment and how they handled the sensitive item compared to the nonsensitive items. It was also tested whether the respondents noticed the sensitive item and whether they were able to distinguish between the sensitive and nonsensitive items. However, in order to answer the question whether the respondents shifted their response patterns (the shifted item difficulty), two online experimental studies were carried out. Study 2 (see Table 5 for an overview) focused on the shifting of item difficulty or rather whether the sensitive item affected or changed the agreement to the nonsensitive items.

This was analyzed by manipulating the position of the sensitive item within the questionnaire, which consisted of three different segments that included sensitive and nonsensitive items presented as direct self-report questions. For this purpose, each item was placed on separate page. A further related question to the inconsistent results of the list experiment is whether the procedure to count the number of items (the number of items that respondents agree to) is distorted. In order to test this assumption, respondents were asked to indicate the number of *yes* answers after two segments, and these values were then compared to the sum of *yes* answers in the direct-self report items. The last study is a replication of Study 2. The hypotheses are tested again in a slightly varied and further developed design. Therefore, the segments were changed. The segment that was presented at the beginning in Study 2 was placed at the end of the questionnaire and vice versa.

Table 5 Main characteristics of Study 2 and Study 3

Study 2			
Condition	**Segment 1**	**Segment 2 (distractor)**	**Segment 3**
Experimental (IV)	One *sensitive* item; Four nonsensitive items (DV); Indication of the number of *yes* answers (DV)	One sensitive item; Four nonsensitive items	One *nonsensitive* item; Four nonsensitive items (DV); Indication of the number of *yes* answers (DV)
Control (IV)	Four nonsensitive items; One *sensitive* item (DV); Indication of the number of *yes* answers (DV)	Four nonsensitive items; One sensitive item	Four nonsensitive items; One *nonsensitive* item (DV); Indication of the number of *yes* answers (DV)

Study 3

Condition	**Segment 1**	**Segment 2 (distractor)**	**Segment 3**
Experimental (IV)	One *nonsensitive* item; Four nonsensitive items (DV); Indication of the number of *yes* answers (DV)	One sensitive item; Four nonsensitive items	One *sensitive* item; Four nonsensitive items (DV); Indication of the number of *yes* answers (DV)
Control (IV)	Four nonsensitive items; One *nonsensitive* item (DV); Indication of the number of *yes* answers (DV)	Four nonsensitive items; One sensitive item	Four nonsensitive items; One *sensitive* item (DV); Indication of the number of *yes* answers (DV)

Notes: Each item was presented on a separate page. In the experimental condition the sensitive and one of the two possible nonsensitive items was presented prior to the four nonsensitive items. IV = independent variable, DV = dependent variable

5. References

Ahart, A. M., & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods, 7*, 101–114.

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly, 58,* 210–240.

Aquilino, W. S. (1997). Privacy effects on self-reported drug use: Interactions with survey mode and respondent characteristics. In L. Harrison & A. Hughes (Eds.), *National Institute on Drug Abuse research monograph series: The validity of self-reported drug use. Improving the accuracy of survey estimates* (No. 167, pp. 383–415). Washington, DC: U.S. Department of Health and Human Services, National Institutes of Health.

Aronson, E. Akert, R. M., & Wilson, T. D. (2004). *Sozialpsychologie [Social psychology]*. München: Pearson Studium (Education).

Assael, H., & Keon, J. (1982). Nonsampling error vs. sampling error in survey research. *Journal of Marketing, 46*, 114–123.

Bauman, K., & Dent, C. (1982). Influence of an objective measure on self-reports of behavior. *Journal of Applied Psychology, 67,* 623–628.

Barton, A. H. (1958). Asking the Embarrassing Question. *Public Opinion Quarterly, 22*, 67–68.

Biemer, P., & Brown, G. (2005). Model-based estimation of drug use prevalence using item count data. *Journal of Official Statistics, 21,* 287–308.

Biemer, P. P., Jordan, B. K., Hubbard, M., & Wright, D. (2005). A test of the item count methodology for estimating cocaine use prevalence. In J. Kennet and J. Gfroerer (Eds.), *Evaluating and improving methods used in the national survey on drug use and health (DHHS Publication No. SMA 05-4044, Methodology Series M-5)* (pp. 149-174). Rockville, MD: Dept. of Health and Human Services Administration, Office of Applied Studies.

Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis, 20*, 47–77.

Blumer, H. (1969). *Symbolic interactionism. Perspective and method*. New Jersey: Englewood Cliffs.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research, 40,* 169–193.

Crowne, D., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349–354.

Dalton, D. R., Daily, C. M., & Wimbush, J. C. (1997). Collecting "sensitive" data in business ethics research: A case for the unmatched count technique (UCT). *Journal of Business Ethics, 16*, 1049–1057.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology*, *47*, 817–828

De Leeuw, E.D (2001) Reducing missing data in surveys: An overview of methods. *Quality & Quantity, 35,* 147–160.

DeMaio, T.J. (1984). Social desirability and survey measurement: A review. In: C.F. Turner, & E. Martin (Eds.), *Surveying subjective phenomena (*pp. 257–281). New York: Russel Sage.

Dickson-Swift, V., James, E. L., & Liamputtong, P. (2008). *Undertaking sensitive research in the health and social sciences: Managing boundaries, emotions and risks.* Cambridge, England: Cambridge University Press.

Dovidio, J. F., & Fazio, R. H. (1992). New technologies for the direct and indirect assessment of attitudes. In J. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 204–237). New York: Russell Sage Foundation.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 185–210). New York: Wiley.

Edwards, A.L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

Esser, H. (1975). *Soziale Regelmäßigkeiten des Befragtenverhaltens* [*Social regularities of response behavior]*. Meisenheim am Glan: Anton Hain.

Esser, H. (1985). Befragtenverhalten als „rationales Handeln“: Zur Erklärung von Antwortverzerrungen in Interviews [Response behavior as „rational action“: To explain response biases in interviews]. *ZUMA-Arbeitsbericht, 85/01*.

Esser, H. (1986). Können Befragte Lügen? Zum Konzept des „wahren Wertes“ im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung. [Can respondents lie? To the concept of the "true value" in the context of the action-theoretical

explanation of situational influences in surveys]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 38*, 314-336.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage Publications.

Groves, R. M., Fowler, F.J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly, 77*, 159–172.

Heitmeyer, W. (2002). *Deutsche Zustände, Folge 1.[German States, Vol. 1]*. Frankfurt am Main: Suhrkamp.

Himmelfarb, S., & Lickteig, C. (1982). Social desirability and the randomized-response technique. *Journal of Personality and Social Psychology, 43,* 710–17.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67,* 79–125.

Holbrook, A.L., & Krosnick, J.A. (2010a). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly, 74,* 328–343.

Holbrook, A. L., & Krosnick, J. A. (2010b). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly, 74*, 37–67.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30,* 161–172.

Janus, A. L. (2010). The influence of social desirability pressures on expressed immigration attitudes. *Social Science Quarterly, 91*, 928–946.

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. In J. Suls (Ed.), *Psychological perspective on the self* (pp. 231–263). Hillsdale, NJ: Erlbaum.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76,* 349–364.

Juster, T., & Smith, J. P. (1997). Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association, 92,* 1268–1278.

Kane, J. G., Craig, S. C., & Wald, K. D. (2004). Religion and presidential politics in Florida: A list experiment. *Social Science Quarterly, 85*, 281–293.

Karlan, D., & Zinman, J. (2012). List randomization for sensitive behavior: An application for measuring use of loan proceeds. *Journal of Developments Economics, 98*, 71–75.

Kolditz, t. A., & Arkin, R. M. (1982). An impression management interpretation of the self-handicapping strategy. *Journal of Personality and Social Psychology, 43*, 492–502.

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research*, *41*, 1387–1403.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity*, *47*, 2025–2047.

Krysan, M., & Couper, M.P. (2003). Race in the live and the virtual interview: Racial deference, social desirability, and activation effects in attitude surveys. *Social Psychology Quarterly, 66*, 364–383.

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial attitudes and the "New South". *The Journal of Politics, 59*, 323–349.

Kuklinski, J. H., Sniderman, P. M., Knight, K., Piazza, T., Tetlock, P. E., Lawrence, G. R., & Mellers, B. (1997). Racial prejudice and attitudes toward affirmative action. *American Journal of Political Science, 41*, 402–419.

LaBrie, J. W., & Earleywine, M. (2000). Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched‐count technique. *Journal of Sex Research, 37*, 321–326.

Lee, R. M. (1993). *Doing research on sensitive topics.* London: Sage.

Lee, R. M., & Renzetti, C. M. (1990). The problems of researching sensitive topics: An overview and introduction. *American Behavioral Scientist, 33*, 510– 528.

Lensvelt-Mulders, G. J. L. M., Hox, J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research. *Sociological Methods and Research, 33,* 319–348.

Lensvelt-Mulders, G. J. L. M., van der Heijden, P. G. M., Laudy, O., & van Gils, G. (2006). A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. *Journal of the Royal Statistical Society Series A, 169*, 305–18.

Locander, W. B., Sudman, S., & Bradburn, N. M. (1976). An investigation of interview method, threat, and response distortion. *Journal of the American Statistical Association, 71,* 269–275.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior.* Unpublished doctoral dissertation, George Washington University.

Mummendey, H. D. (2006). Selbstdarstellung (self-presentation). In H. W. Bierhoff & D. Frey (Eds.), Handbuch der Sozialpsychologie und Kommunikationspsychologie (pp. 49–56). Göttingen: Hogrefe.

Mummendey, H. D., & Bolten, H.G. (1985). Die Impression-Management-Theorie. In D. Frey & M. Irle (Eds.), *Theorien der Sozialpsychologie. Band III: Motivations- und Informationsverarbeitungstheorien [Theories of social psychology. Volume III: Theories of motivation and information processing]* (pp. 57–77). Bern: Huber.

Murray, D., O'Connell, C., Schmid, L., & Perry, C. (1987). The validity of smoking self-reports by adolescents: A reexamination of the bogus pipeline procedure. *Addictive Behaviors, 12,* 7–15.

Musch, J., Brockhaus, R., & Bröder, A. (2002). Ein Inventar zu Erfassung von zwei Faktoren sozialer Erwünschtheit. *Diagnostica, 48*, 121–129.

Näher, A.-F., & Krumpal, I. (2012). Asking sensitive questions: The impact of forgiving wording and question context on social desirability bias. *Quality and Quantity, 46*, 1601–1616.

Ostapczuk, M. S. (2008). *Experimentelle Umfrageforschung mit der Randomized-Response-Technik [Experimental survey research with randomized response technique]*. Doctoral dissertation, Düsseldorf: Heinrich Heine University.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609.

Phillips, D. L., & Clancy, K. J. (1972). Some effects of "social desirability" in survey studies. *American Journal of Sociology, 77,* 921–940.

Phillips, D.L. (1971). *Knowledge from what?* Chicago: Rand McNally.

Raghavarao, D., & Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society Series B (Methodological), 41*, 40–45.

Randall, D. M., & Fernandes, F. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics, 10,* 805-817.

Rasinski, K.A., Baldwin, A.K., Willis, G.B., Jobe, J.B. (1994*). Risk and loss perceptions associated with survey reporting of sensitive topics.* National Opinion Research Center (NORC), Chicago, 497–502.

Rasinski, K.A., Willis, G.B., Baldwin, A.K., Yeh, W.C., Lee, L. (1999) Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology, 13,* 465–484.

Rayburn, N. R., Earleywine, M., & Davison, G. C. (2003a). Base rates of hate crime victimization among college students. *Journal of Interpersonal Violence, 18,* 1209–1211.

Rayburn, N. R., Earleywine, M., & Davison, G. C. (2003b). An investigation of base rates of anti-gay hate crimes using the unmatched-count technique. *Journal of Aggression, Maltreatment & Trauma, 6,* 137–152.

Reinecke, J. (1991). *Interviewer- und Befragtenverhalten: Theoretische Ansätze und methodische Konzepte [Interviewer- and respondents behavior: Theoretical approaches and methodological concepts]*. Opladen: Westdeutscher Verlag.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin, 114,* 363–375.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Schlenker, B.R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations.* Belmont, CA: Brooks/Cole.

Schuman, H., & Converse, J.M. (1971). Effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly, 35*, 44–68.

Schnell, R., Hill, P. B., & Esser, E. (2005). *Methoden der empirischen Sozialforschung [Methods of emprical social research]*. München: Oldenbourg.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Sieber, J. E., & Stanley, B. (1988). Ethical and professional dimensions of socially sensitive research. *American Psychologist, 43*, 49–55.

Smith, T. W. (1992). Discrepancies between men and women in reporting number of sexual partners: A summary from four countries. *Biodemography and Social Biology, 39,* 203–211.

Sudman, S., & Bradburn, N. (1974). *Response effects in surveys: A review and synthesis.* Chicago: Aldine.

Sudman, S., & Bradburn, N. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Stocké, Volker (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Rational-Choice Theorie und des Modells der Frame-Selektion [Determinants for respondents' susceptibility to social desirability bias: A

comparison of predictions from rational choice theory and the model of frame-selection]. *Zeitschrift für Soziologie, 33,* 303–320.

Stocké, V. (2007a). Determinants and consequences of survey respondents' social desirability beliefs about racial attitudes. *Methodology, 3*, 125–138.

Stocké, V. (2007b). The interdependence of determinants for the strength and direction of social desirability bias in racial attitude surveys. *Journal of Official Statistics, 23,* 493–514.

Streb, M. J., Burrell, B., Frederick, B., & Genovese, M. A. (2008). Social desirability effects and support for a female American president. *Public Opinion Quarterly, 72*, 76–89.

Stulginskas, J. V., Verreault, R., & Pless, I., B. (1985). A comparison of observed and reported restraint use by children and adults. *Accident Analysis & Prevention, 17,* 381–386.

Tedeschi, J. T., Lindskold, S., & Rosenfeld, P. (1985). *Introduction to social psychology*. St. Paul: West.

Tedeschi, J. T., & Riess, M. (1981). Identities, the phenomenal self, and laboratory research. In J.T. Tedeschi (Ed.), *Impression management theory and social psychological research,*(pp.3–22). New York: Academic Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859–883.

Tourangeau, R., Smith, T. W., & Rasinski, K. (1997). Motivation to report sensitive behaviors in surveys: Evidence from a bogus pipeline experiment. *Journal of Applied Social Psychology, 27,* 209–222.

Tsuchiya, T., Hirai, Y., & Ono, S. (2007). A study of the properties of the item count technique. *Public Opinion Quarterly, 71*, 253–272.

Volicer, B. J., & Volicer, L. (1982). Randomized response technique for estimating alcohol use and non compliance in hypertensives. *Journal of Studies on Alcohol*, *43,* 739-750.

Warner, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60,* 63–69.

Willis, G.B., Sirken, M., Nathan, G. (1994). *The cognitive aspects of responses to sensitive survey questions.* (Working Paper Series 9.). Hyattsville, MD: National Center for Health Statistics, Cognitive Methods Staff.

Wimbush, J. C., & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of the Applied Psychology, 82*, 756–763.

Winkler, N., Kroh, M., & Spiess, M. (2006). Entwicklung einer deutschen Kurzskala zur zweidimensionalen Messung von sozialer Erwünschtheit [Development of a German short scale for two dimensional measurement of social desirability]. *German Institute for Economic Research. Discussion Papers*, *579*.

Wolf, F. (2012). *Heikle Fragen in Interviews: Eine Validierung der Randomized Response-Technik [Delicate questions in interviews: A validation of the randomized response technique]*. Wiesbaden: Springer VS.

Wyner, G. A. (1980). Response errors in self-reported number of arrests. *Sociological Methods and Research, 9,* 161–177.

Zick, A., Wolf, C., Küpper, B., Davidov, E., Schmidt, P., & Heitmeyer, W. (2008). The syndrome of group-focused enmity: The interrelation of prejudices tested with multiple cross-sectional and panel data. *Journal of Social Issues, 64*, 363–383.

Zimmerman, R. S., & Langer, L. M. (1995). Improving estimates of prevalence rates of sensitive behaviors: The randomized lists technique and consideration of self-reported honesty. *Journal of Sex Research, 32*, 107–117

Manuscript #1:

Is the List Experiment doing its Job?

Inconclusive Evidence!

Stefanie Gosen[1], Peter Schmidt[2], Stefan Thörner[1], & Jürgen Leibold[3]

[1]Phillipps-University Marburg, Germany

[2]Justus-Liebig-University Gießen, Germany and National Research University Higher School of Economics, Moscow, Russia

[3]Georg-August-University Göttingen, Germany

Submission date: October 08, 2013

Under review in *Journal of Official Statistics*

Abstract

This paper sheds new light on the unobtrusive measure known as the 'list experiment' or 'unmatched count technique'. Proponents of this method claim that it detects social desirability bias in responses to sensitive questions in surveys. The logic of this method is quite straightforward. After a critical overview of the theory, logic, and empirical results of this type of measure, we present the results of a series of three studies. While the first study yielded promising results, the replication of the outcome pattern in study 2 failed completely. The third study, based on longitudinal data, delivers indications for the systematic inconsistency of the claimed logic of the 'list experiment'. First, significant mean differences between baseline and test condition occur even if the additional item is nonsensitive and has an agreement rate of about two percent in direct questioning. Second, test-retest reliability shows fluctuating results depending on the sensitivity and number of items included in the experiment. Implications for theory and practice in measuring social desirability by unobtrusive measures are discussed.

Keywords: social desirability; indirect survey techniques; sensitive question; Islamophobia; anti-Semitism

1. Introduction

The measurement of sensitive topics has been a constant challenge in social science research. The notion of sensitive questions presupposes that respondents believe there are norms defining desirable attitudes and behavior patterns. Furthermore, one may assume that they are concerned enough about these norms that they distort their answers in order to avoid presenting themselves in an unfavorable light. Misreporting gets worse as the topic becomes more sensitive among those with something to hide, and it seems to be reduced if self-administered surveys are administered and techniques such as randomized response are used (Tourangeau, Rips and Rasinski 2000, p. 257). When participants provide socially desirable responses to sensitive questions, it could lead to highly biased descriptive statistics (i.e., means and standard deviations). An example of this can be found in the assessed prevalence rates of behaviors like abortion, condom use, or deviant behavior (see, e.g., Nickel et al. 1995; Tourangeau and Yan 2007) or attitudes toward, for example, minorities (see, e.g., Huddy and Feldman 2009; Salzborn 2010). In addition, sensitive questions may lead to biased parameters of correlations and regression coefficients. This was demonstrated, for instance, by Oberski, Weber and Révilla (2012) in an experimental study using data from the European Social Survey, which demonstrates how the coefficients in a structural equation model may be biased due to social desirability. Furthermore, it may lead to an increase in the percentage of missing values, which are not completely at random. This refers to variables like income but also to attitudes toward minorities or deviant behavior. Such missing value patterns may also result in biased estimates of means, standard deviations, correlation, and regression coefficients (Papastefanou and Wiedenbeck 1998).

Even though researchers and social science research textbooks do not ignore the problem of the social desirability response set, the solutions that are offered for overcoming this and the related empirical evidence are not very clear on how to deal with this problem.

In practice, there are at least two different ways of controlling the social desirability response bias. On the one hand there are methods that allow an individual measurement of the bias, like social desirability scales (Crowne and Marlow 1960; Paulhus and Reid 1991). On the other hand there are methods which demonstrate the existence of the social desirability bias without controlling the individual extent of the bias, for example, the list experiment (Kuklinski, Cobb and Gilens 1997; Kuklinski, Sniderman et al. 1997) or the bogus pipeline effect (Jones and Sigall 1971). There is neither a systematic comparison of the different approaches such as the measurement as an individual attribute (Paulhus and Reid 1991; Schuessler 1982) or the list experiment (however, an overview is given by Krumpal 2011) nor consistent empirical evidence available through a series of systematic meta-analyses of the effects of controlling for social desirability (SD) response sets by either SD scales or unobtrusive measures (Sniderman and Grob 1996). To our knowledge there is only one meta-analysis available by Tourangeau and Yan (2007), which is based on seven studies only, and another meta-analysis in preparation (Auspurg et al. 2012). Therefore, researchers have no solid information concerning which instruments might be best suited to control for social desirability to make evidence-based decisions for planning their projects and research designs. This may be a reason why, until now, large data-generating programs such as the European Social Survey (ESS), the General Social Survey (GSS) in the United States, the International Social Survey Program (ISSP), the World Value Study (WVS), and the European Value Study have not implemented any instrument for the continuous and systematic control of social desirability effects. All these comprehensive data-generating programs have neither used social desirability scales nor indirect or unobtrusive measures like the list experiment.

Social desirability response bias arises especially in self-reports. This occurs when the question has a potentially embarrassing, stigmatizing, or incriminating character (Dalton,

Wimbush and Daily 1994, p. 817). By using unobtrusive measures, the respondents should be guaranteed anonymity specifically for sensitive topics such as voting, attitudes toward minorities, income, sexual behavior, etc. (Chaudhuri and Christofides 2007; Holbrook and Krosnick 2010; Janus 2010; Kuklinski, Cobb and Gilens 1997; Kuklinski, Sniderman et al. 1997; LaBrie and Earleywine 2000). In such cases, it is assumed that the respondents have no motivation to present themselves in a socially acceptable way, because of the assured anonymity. As a consequence, it is postulated that the respondent answers the questions more honestly and truthfully (Warner 1965; Himmelfarb and Lickteig 1982; Paulhus 1984).

One of the most well-known indirect/unobtrusive survey methods for controlling the social desirability response set is the "randomized response technique" (RRT) which originates from Warner (1965) (see Lensvelt-Mulders, Hox and van der Heijden 2005). An additional method, which has been gaining more popularity, is the list experiment (Sniderman and Grob 1996; Kuklinski, Cobb and Gilens 1997; Kuklinski, Sniderman et al. 1997). The list experiment was first conceptualized by Miller (1984); he called his new contribution the „item count technique". It is also known as "unmatched count technique" (Dalton, Wimbush and Daily 1994), "randomized list technique" (Zimmerman and Langer 1995), and "list randomization technique" (Karlan and Zinman 2011). Another form which has a high degree of similarity with the list experiment is the "block total response method" developed by Raghavarao and Federer (1979).

Thus, this paper is structured as follows: After an initial discussion on the basic rationale of the list experiment, an overview of various studies using this technique will be given. Next, we introduce and describe our three studies examining this methodology. Finally, we present a summary of the main results and discuss the possible inferences that can be made based on our findings.

The List Experiment

The underlying idea of the list experiment is to provide the respondents with a feeling of anonymity and privacy when answering questions. Its basic premise is that respondents will answer in a more honest rather than in a socially desirable way. The logic of the method is as follows: The respondents are divided randomly into two groups. One half of the respondents (baseline condition) receive a list with, for example, three statements. These three statements are nonsensitive questions and should be neutral in character. In the next step, the respondents should indicate how many of the items make them angry. In this step it is important that the respondents understand that they only have to specify the number of the items and not the specific items which make them angry. An example of a list experiment in the domain of racism is given by Kuklinski, Sniderman et al. (1997, p. 405):

"Now I`m going to read you three things that sometimes make people angry or upset. After I read all three, just tell me how many of them upset you. I don´t want to know which ones, just how many".

1. *"the federal government increases the tax on gasoline;"*
2. *"professional athletes getting million dollar salaries;"*
3. *"large corporations polluting the environment."*

The other half of the respondents (test condition) were presented the same three basic statements plus an additional statement. This additional item contains the sensitive topic that is under study. In Kuklinski, Sniderman et al.'s (1997) study, the sensitive item number four was *"a black family moving in next door"*. Respondents are then asked to indicate the number of items which made them angry.

By using this procedure it is possible to generate a proportion of those respondents who are angry about the sensitive statement at the aggregate level (Dalton, Wimbush and

Daily 1994; Kuklinski, Sniderman et al. 1997; Ahart and Sackett 2004; Holbrook and Krosnick 2010). This works as follows: Firstly, one determines the mean levels of the items that are considered to cause anger in both groups. Secondly, the mean level of the test condition is subtracted from the mean level of the baseline condition. To compute the percentage, the mean difference of the two groups is taken *100. The underlying logic of the list experiment implies that every increase in the mean of those statements which made the respondents angry in the test condition must be attributed to the sensitive statement. This means, for example, if the baseline condition has a mean value of about 2.5 "angry" statements and the mean in the test condition is 3.0, that half of the respondents of the test condition are angered by the sensitive statement (Streb et al. 2008, p. 81).

There are numerous studies which have applied the list experiment, and they all have shown that it was used successfully to identify and correct for sensitive questions (Miller 1984; Dalton, Wimbush, Daily 1994; Kuklinski, Cobb and Gilens, 1997; Kuklinski, Sniderman et al. 1997; Kane, Craig and Wald 2004; Streb et al., 2008; Coutts and Jann 2011; Blair and Imai 2012). However, Zigerell (2011) reported evidence that some respondents tend to deflate the reported number of items in the list experiment. These people try to avoid being associated with a socially undesirable item. Furthermore, Holbrook and Krosnick (2010) found a difference between survey methods (mode effect) and showed that the list experiment did not have higher estimates in Internet surveys. In addition, Glynn concluded that the one of the major disadvantages of the list experiment is the necessity of a large sample size. Furthermore, a standard analysis does not allow checking for the many possible violations of behavioral assumptions which Glynn refers to. Finally, according to Glynn, it is difficult to use the standard procedure of a list experiment in multivariate models. Concerning the last point, however, there are procedures now available for multivariate techniques (Imai 2011).

In the present paper, in a sequence of three studies (see table 1) we combine the following attributes, which represent the added value of the paper. By replicating the list experiment we test the invariance of the results across different samples (see study 1, study 2, and study 3). Specifically,

1. We conduct a series of interrelated, consecutive studies to test the external validity of the findings.
2. We use different data collection methods to take mode effects into account. We employ CATI (Computer Assisted Telephone Interviews) in studies 1 and 2 and online survey data in study 3.
3. The application of a panel design allows testing the temporal stability in study 3.
4. The sensitive questions utilize depict current issues of concern in many societies (Islamophobia and anti-Semitism).

The special feature of our paper is the uniqueness of the studies. Different variants of list experiments were combined, confronted, and compared.

2. Design, Sample, Method, and Results

For testing our propositions and testing the validity of the list experiment methodology, we conducted three different studies implementing a total of five list experiments. Study 1 is based on a representative sample of the German voting population and is part of the project "Group-Focused Enmity" (GFE) sponsored by the Volkswagen Foundation (for a precise sample description see, e.g., Heitmeyer 2002; Zick et al. 2008,). The data were collected between the end of April and mid-June 2009 with the CATI (Computer Assisted Telephone Interview) method via a professional survey institute. The sample size is N = 229 and the mean age of respondents is 52.15 years. The list experiment was conducted with three items in the baseline condition and four items in the test condition.

Study 2 is also based on a representative sample of the German voting population and is also part of the GFE project, sponsored by the Volkswagen Foundation (for a precise sample description see, e.g., Heitmeyer 2002; Zick et al. 2008). The data were collected between the end of April and mid-June 2010 with the CATI method by a professional survey institute. The sample size is N = 445 and the mean age of respondents is 53.88 years. In this study, the list experiment was conducted with four baseline items and five items in the test condition.

Study 3 is a panel study of a non-representative online survey. The first wave was collected in January 2011. The sample size is N = 1,569. The mean age of respondents is 27.62 years. Regarding education, 58.9 percent of the respondents had completed high school and 32.4 percent have a university degree. Furthermore, 61.1 percent of participants are female and 38.9 percent male.

To check the validity of the list experiment, we tested four different types of list experiments. We conducted two baseline conditions, one with four baseline items and another one with the same four baseline items plus a fifth nonsensitive (neutral) item. In addition, we created two different test conditions, each with five statements. The second wave was collected in June 2011. A total of 194 respondents completed the survey[1]. Even if the second wave is relatively small compared to the first one, we cannot find any systematic attrition effects. Therefore, wave 2 can be regarded as an appropriate subsample of wave 1. In this wave we also used the same baseline and test conditions as in wave 1. The availability of panel data allows us to check the reliability of the list experiment using test-retest correlations.

[1] In wave 1 of study 3 we asked respondents for their willingness to participate again in a similar study. We gathered, on a voluntary basis, around 500 email addresses which we used to conduct the second wave (N = 194 completed the second questionnaire).

[Table 1 about here.]

Study 1

The main focus of this study is the list experiment itself and its implementation in a survey examining prejudice toward Jewish people in Germany. To conduct the list experiment, the study´s participants were randomly split into two groups. Following Kuklinski, Sniderman et al. (1997), in our first study we used three items in the baseline and four items in the test condition.

The question wording for the baseline condition (N = 108) was as follows:

"Now I will refer to some topics people occasionally express anger[2] about. Could you please tell me how many of the following three statements have also made you angry.[3]

1. *The way gas prices keep going up,*
2. *That professional athletes get million dollar salaries,*
3. *That the German railroad has so many delays."[4]*

"Have you ever been angry about three, two, one, or none of the statements mentioned?"

[2] Anger in German is Wut/ wütend; Ärger/ sich ärgern.
[3] The term "angry" in the introduction of the list experiment was taken from Kuklinski, Sniderman et al. (1997). They assume that anger implies salience. The more the statements make people angry, especially the sensitive statement, the more meaningful the items are for them. They have pointed out that it is possible that anger motivates behavior more than any other negative emotions (Kuklinski, Sniderman et al. 1997, pp. 413-414).
[4] Statement 1 was taken from Streb et al. (2008). We used this statement with the same wording. However, due to the permanent increase of the "gasoline prices" in Germany, we had to change the word "gasoline" to "gas". The risk would be too high that too many respondents would get angry about the statement. Therefore, a ceiling effect cannot be controlled. Statement 2 was taken from Kuklinski, Sniderman et al. (1997). Both statements were used in translated versions. We developed statement 3 to present a topic which is specific for Germany (German railroad – Deutsche Bahn).

Participants in the test condition (N = 121) received these three basic statements along with a fourth, additional one. This statement contains the sensitive topic which is anti-Semitism. The wording is as follows:

4. *"That Jews have too much influence in the world."*[5]

Accordingly, the question is:

"Have you ever been angry about four, three, two, one, or none of the statements mentioned?"

The aim of this survey is to ascertain prejudices toward Jews, which are probably inhibited in the attitude items of the GFE syndrome (Zick et. al. 2008) because of social desirability bias. Based on earlier research conducted by Streb et al. (2008), Kuklinski, Sniderman et al. (1997), and Gilens, Sniderman, and Kuklinski (1998), for instance, the central assumption is that socially desirable responding has an effect on nonanonymous items. Both Streb et al. (2008) and Kuklinski et al. (1997) argue that the list experiment, by its design, should be not biased by a social desirability response set.

[Table 2 about here.]

Table 2 presents the mean levels of feeling anger in the baseline and in the test conditions. Furthermore, it shows the percentage of people who reported having gotten angry about the statement "that Jews have too much influence in the world". One can see from table 2 that there is a significant difference of 48.62 percent between baseline and test conditions. By following the logic of the interpretation of the list experiment, one can infer that 48.62 percent of the respondents are angry about the influence of Jews in the world. As Kuklinski,

[5] This item is part of the battery to measure anti-Semitism in the GFE project (Zick et al. 2008; Davidov et al. 2011).

Sniderman et al. (1997), we assume that respondents are angry about the substantial issue and not about the statement itself.

The direct question to investigate social desirability response bias is as follows: "Jews have too much influence in Germany". This item is a personal attitude statement and was assessed with a four-point Likert scale (1 fully agree – 4 fully disagree). Unfortunately, we did not have the same item as in the list experiment[6], so we had to use a similar item for this comparison. We know that not having the same items can be a problem. However, we had no other possibility for a comparison. In the next two studies we corrected this issue by using the same items.

If one follows the predominant view of the existing literature (Droitcour et al. 1991; Dalton, Wimbush and Daily 1994; Kuklinski, Cobb and Gilens 1997; Kuklinski, Sniderman et al. 1997), study 1 shows evidence of a social desirability response bias: 11.7 percent of the respondents agree or rather agree to the direct self-report question "Jews have too much influence in Germany". In contrast, within the list experiment, 48.62 percent of respondents can be assumed to hold anti-Semitic beliefs. By following the logic of the list experiment, this is an indication that respondents might not have stated their true opinion about Jews while answering the common, low privacy, and low anonymity direct question. According to the literature, we can assume that people have the tendency to present themselves as unprejudiced (e.g., Dovidio and Gaertner, 1986). To check for significant differences between the direct self-report item and the list experiment, we used the "Difference of Proportion Test."[7] The list experiment, as a high anonymity condition, revealed a higher confirmation rate regarding anti-Semitism ($z = 2.52$, $p < .01$).

[6] "Jews have too much influence in the world."

[7] Test statistic : $z = (abs(p1\text{-}p2))/(sqrt(SE_1^2+SE_2^2))$. Standard error of a proportion of a direct self-report: SE_p = Square Root $((p*(1-p))/n)$. Standard error of a proportion of list experiment was the standard error of the difference. (see Holbrook and Krosnick 2010, p. 51)

[Table 3 about here.]

A frequency analysis can be done to search for possible ceiling effects[8]. To avoid these effects, the test condition usually receives five or more statements (Kuklinski, Cobb and Gilens 1997; Kuklinski, Sniderman et al. 1997; Streb et al. 2008). In this study, the list experiment uses a maximum of four statements in the test condition; hence, it is possible that a ceiling effect might exist. However, in this case, the frequency analysis reveals low indications for possible ceiling effects. About 24 percent of the respondents of the test condition show anger concerning three statements. If they would express their anger about all four statements they would not be able to conceal their prejudiced attitudes toward Jews. Thus, the high-privacy condition would be violated. The interviewer would know that the respondent felt angry about every item including the sensitive one. In view of this fact, the high-privacy condition of the list experiment is no longer fulfilled. Therefore, social desirability bias cannot be totally excluded in the list experiment.

Study 2

This study is a slightly modified repetition of study 1. The crucial question examined here is whether it would be possible to replicate the results of the previous study.

Given the problems of our first study, we decided to make some changes to obtain more reliable results. In study 1 we used four items in the test condition, and with this number of statements it was not possible to rule out ceiling effects (Kuklinski, Sniderman et al. 1997; Streb et al. 2008). To avoid this effect, we added another item which was expected

[8] Ceiling effects occur when respondents agree to the nonsensitive items as well as to the sensitive statement (Blair and Imai 2012).

to evoke less distress from the respondents. This means that the probability that respondents would report feeling angry about all four items decreased. The new item was:

"That seat belts are fastened when driving".[9]

Secondly, we extended the introduction. Furthermore, we increased the baseline condition from three to four and, accordingly, the test condition from four to five items. Due to this, the respondents were instructed to count the number of items which make them angry or upset with their fingers.[10] We think that this technique would help respondents to not become confused. Specifically, they have the possibility to decide immediately, after the interviewer reads the respective statement, whether or not they agree (in our case: whether they are angry or not). The advantage of this strategy is that the respondents do not have to keep all four or five statements in mind before giving their answer in the second part of the question.

The last and third improvement concerns the direct self-report item for the control condition. In study 1 we did not use the same wording for this item as was used in the list experiment. Hence, a comparison was problematic. To obtain an adequate reference score, only the baseline condition included the direct self-report item "Jews have too much influence in the world", which was set in the same block with the GFE syndrome items. The reason for posing the direct question only in the baseline condition was that the test condition would otherwise receive the stimulus twice. Firstly, it would then be posed once with the self-report attitude items of the GFE syndrome and then a second time in the list experiment. The possibility is therefore high that the respondents could notice the item in the first round

[9] The original item is taken from Streb et al. (2008).
[10] The new introduction was: "Now I will refer to some topics people occasionally express anger about. Could you please tell me how many of the following four (five) statements you have also been angry about. Please count, using your fingers, how many you have been angered by."

and, as a consequence, this increases the risk that the respondents might foresee the intention of the list experiment thus resulting in substantial bias.

The mean levels of anger for both conditions (baseline and test condition) are shown in table 4. In study 2, the difference between baseline and test condition amounts to 9.48 percent (ns) in the expected direction. The self-report attitude item "Jews have too much influence in the world" was answered affirmatively by 24.1 percent of respondents (see table 3, $z = 1.41$, $p < .10$). The difference between indirect and direct question is not significant at the 5% level but it is at the 10% level. This result indicates that almost twice as many respondents agree to the direct self-report question of the syndrome compared to the unobtrusive measure. Finally, this finding shows that the list experiment provides no evidence of detecting social desirability response bias in this study.

[Table 4 about here.]

In addition, a frequency analysis of the list experiment showed that the majority of respondents from the test condition expressed anger concerning two or three statements. Only 6.4 percent expressed anger about four of five[11] statements. This minimal percentage allows the exclusion of a possible ceiling effect.

However, it is questionable how the inconsistent results and the resulting doubtful validity appeared so profoundly in this context. One possible explanation might be that a sequence effect was operating in this case (Schwarz and Sudman 1996). While the list experiment was performed in study 1 before the self-report attitude items of the syndrome, it was presented directly after the attitude items in study 2.

[11] Anger about all five statements is expressed by1.4 percent of the respondents.

The sequence was as follows:

Study 1 → list experiment before anti-Semitism items

Study 2 → anti-Semitism items before list experiment

An additional reason for the inconsistent results of the list experiment, however, might be the different numbers of statements in varying conditions. In other words, we hypothesize that the agreement to the statements does not only depend on content. The sole number of the listed statements leads to more "yes" answers. In this regard, the increased mean in the test condition is not necessarily caused by the sensitive item that should be examined. In this case it is the item "that Jews have too much influence in the world". In study 3 we consider this phenomenon more precisely by performing a more elaborate investigation comparing four different types of list experiments.

Study 3 (Panel)

Wave 1

Four list experiments with various numbers of statements and with two different sensitive statements were conducted. We first start with the description of the four list experiments. To avoid any sequence effects, the list experiments were presented continuously at the beginning of the questionnaire. Furthermore, the allocation of the list experiment to the respondents was done with a random generator embedded in a weblink.[12] Additionally, the statements of the list experiment were rotated in this survey. The reason for the use of the rotation procedure can be seen in possible position effects of the sensitive statement.

The introduction for the four list experiments was the same, only parts of the statements were changed. Changes can be seen in the following sections. The first

[12] Randomization was successful. There was no systematic difference between groups concerning sociodemographic variables.

experimental group completed the list experiment (N = 376) that included only four nonsensitive statements[13]:

"Below you can see a list of some topics that people express anger about occasionally. How many of these things make or made you angry:

1. *That the costs of gasoline rise regularly at Easter time and during the school holidays.*
2. *That professional athletes are getting million dollar salaries,*
3. *That the German railroad has so many delays,*[14]
4. *That seat belts are fastened when driving."*[15]

How many of these topics made you feel upset?

0, 1, 2, 3, or 4.

The list experiment presented to the second experimental group (N = 379) corresponds to the design of the experiments from the first two studies to get a comparable score. The sensitive and prejudiced statement is:

"That Jews have too much influence in the world."[16]

A further step to test the validity was to execute the experiment with a different item from the syndrome battery of GFE. To avoid possible distortions that might have emerged from the previously analyzed prejudices toward Jews, we decided to include a politically charged issue. The new items chosen represent the construct Islamophobia, which is also part of the GFE syndrome (see Zick et al. 2008). To obtain accurate comparability to anti-Semitism, the content of the statements was the same with the exception of the word "Jews"

[13] In studies 1 and 2 this is called baseline condition.
[14] Statement 2 was taken from Kuklinski, Sniderman et al. (1997) in a translated version. We developed statements 1 and 3 to present special German topics.
[15] The original item was provided by Streb et al. (2008).
[16] The other statements remained constant.

being replaced with the word "Muslims". Therefore, the statement in the third condition (N = 404) is:

"That Muslims have too much influence in the world."[17]

As mentioned above, we suppose that the change or increase of the mean in the test condition might be caused by the higher number of statements. For this reason we developed a new statement. Thereby, we were particularly careful that this statement was very neutral or nonsensitive so that the majority of the respondents would not be angered by it. Instead of a sensitive item that usually receives the respondents´ attention, we presented a nonsensitive statement. This is the only way possible to determine whether respondents make their decision based simply on the number of statements or on the content of the statements. The four "basic" statements of the fourth list experiment condition (N = 410) remained constant. We only added the new, nonsensitive statement:

"That one can easily make purchases besides gas at gas stations if necessary."[18]

[Table 5 about here.]

For an initial overview, table 5 presents the mean levels of anger from all four tested list experiments. Furthermore, table 6 allows the comparison of the four experiments and the examination of the mean differences between each condition.

[17] As in condition 2; the other statements remained constant.

[18] In Germany, by law, supermarkets and shops are closed on Sundays and usually after 10 P.M. on working days. Thus, there is no 24/7 service available in regular shops. Only kiosks and convenience stores at gasoline stations are allowed to provide a 24/7 service.
To test the neutrality of the gasoline station item, we included it in a pretest survey as a direct self-report question. The exact wording was: "I am angry that one can easily make purchases at gas stations if necessary." A frequency analyses showed that only 2 of 75 respondents have been angry about this item.

[Table 6 about here.]

Between the first condition (four statements) and the second one (four statements + statement about Jews), no significant difference was found. This result is similar to what was found in study 2. Thereby, a possible sequence effect, which in study 2 was the presumed reason for the contrasting results of study 1, can be excluded.

However, there is a significant difference of 33.69 percent between the first condition (four statements) and the third condition (four statements + statement about Muslims). Following the logic of the list experiment, this finding reveals to us that 33.69 percent of the respondents express themselves as Islamophobic.

As mentioned above, we also implemented a condition with five nonsensitive items (four statements + nonsensitive statement) only. By considering this condition as a second baseline, we now present the following results:

The difference between conditions four and two was -14 (ns) percent. This result is definitely not in the expected direction. Concerning the content and according to the logic of the list experiment, it means that descriptively more respondents express anger about the possibility of purchases in gas stations than about the influence of Jews in the world. The difference between conditions four and three is nonsignificant (i.e., 11.28%). Nevertheless, the positive value indicates a difference in the expected direction.

The results above appear to be strange. This is due to the rigorous interpretation of the data following the logic of the list experiment. An often mentioned but seldom checked criticism is the comparison of conditions with different numbers of items. The question is, therefore: Is it possible that higher mean values could solely occur because of the higher

number of items? The answer is: Yes, it is possible. Data analysis reveals a significant difference between the first and fourth condition (four statements vs. four statements + nonsensitive statement, respectively) amounting to 22.4 percent. That implies a strong cut concerning the validity of the experiment itself. In other words, no significant mean difference should result if one assumes that the statement "one can easily make purchases at gas stations if necessary" is really neutral. But, in fact, the respondents showed more anger in the four statements + nonsensitive statement condition than in the common baseline condition (four statements).[19] Referring to the substantial interpretation of the data above, this suggests that the respondents express their anger not only in regard to the statement "that Muslims have too much influence in the world", the number of items matters as well. However, it is simply not possible to disentangle the two reasons for the mean increase - item content and higher number of items. Without this knowledge it can be assumed that more than 33.69 percent of the respondents were angry that Muslims have too much influence in the world. The substantial interpretation becomes untenable considering these new results. The increase of the mean cannot be only attributed to the sensitive content of each statement. Rather, it is a mixture of content and number of statements which leads to more "yes" answers.

Furthermore, frequency analyses of the list experiments which include the sensitive statements indicate that no ceiling effects exist. Only 2.6 percent of the respondents express anger about four statements in the second condition (four statements + statement about Jews), while in the third condition (four statements + statement about Muslims), 5.2 percent are

[19] Holbrook and Krosnick (2010) tested the list length in an Internet survey (nonrepresentative) although using a different introduction and other neutral statements. However, they found no differences between the shorter and the longer list.

angry about four statements. Furthermore, only 2.4 percent show anger about four statements in the fourth condition.[20]

In order to allow a comparison of social desirability response bias, as was done in the two previous studies, we measured the direct self-report items in a separate item block in this study (seven-point Likert scale, 1 fully agree – 7 fully disagree). However, only two possible prejudice questions were posed: "Muslims have too much influence in the world" and "Jews have too much influence in the world". Again, only the respondents who were not in one of the respective test conditions in the list experiment received the self-report question.

In the second condition, 13.7 percent of the respondents agree or rather agree to the direct self-report item "Jews have too much influence in the world". However, in the list experiment, only 8.4 percent express themselves to be anti-Semitic. In this second group, no difference within the list experiment is identifiable. Furthermore, the difference between direct question and list experiment is nonsignificant (see table 3, $z = 1.25$, ns) but reveals a tendency that the respondents answered more truthfully to the direct question. This result shows almost the same inconsistency as in study 2. Based on this, one may assume that the list experiment is not able to detect a social desirability response bias when the sensitive statement is about Jews. Therefore, the distorted results from study 2 are not necessarily caused by a sequence effect. Due to the experimental conditions, the underlying causal mechanism was tested for and confirmed. The distorted results are rather a consequence of the not existing validity of the list experiment.

Only on one point does the third condition indicate a more valid result. In contrast to the second condition, the difference of 33.69 percent within the list experiment is in the expected direction. However, 26.8 percent of the respondents agree or rather agree with the

[20] Only 1.3 percent express anger about all five statements in the condition with the anti-Semitic statement, 1.7 percent express anger about five statements in the condition with the Islamophobia statement, and 1.0 percent express anger about all five statements in the fourth condition.

direct self-report Islamophobia item. This result indicates that list experiment respondents who are angry about the item "Muslims have too much influence in the world" do not differ significantly from the respondents in the direct self-report condition (see table 3, z = .92, ns).

In summary, one might assume from this finding that prejudice attitudes toward Muslims can be detected by using the list experiment. However, by the control of the baseline condition with four statements + nonsensitive statement in the second part of the analyses, it is apparent that almost as many of the respondents are angry about one or more neutral or nonsensitive statements. As mentioned above, this means that the respondents express anger not only in regard to the statement "Muslims have too much influence in the world" but also to one or more neutral statements.

Second Wave

The reason for the use and analysis of the data from the second wave was to replicate our findings and to ensure the accuracy of the results from the previous studies. Furthermore, we tested the stability of the aggregate level and the intraindividual stability of the instrument for waves 1 and 2. In the second wave, we used the same four list experiments, making no changes in the instructions and items.

Table 5 indicates the sample size of each condition and the mean level of expressed anger for all tested list experiments. In addition, the mean differences between the differing conditions of list experiments are presented in table 6.

Aggregate level

The mean difference between condition one (four statements) and condition two (four statements + statement about Jews) amounted to 8.81 percent.[21] This result is almost equal to the difference in wave 1. At first glance, no change in the responses is identifiable over time.

[21] We were able to replicate the mean levels and mean differences. Due to the small sample size in wave 2, significance was not reached.

However, in wave 2, the mean difference between the first condition and the third (four statements + statement about Muslims) increases to a value of 40.45 percent. It appears that the responsiveness changes over time and that the rejection of Muslims has increased. Furthermore, the mean difference of 12.65 percent between condition one and four (four statements + nonsensitive statement) indicates a decrease from time one to time two. The difference is not as high as in study 3 (wave 1) and demonstrates a change in regard to how respondents answered the question over time.

We also used the fourth condition as the baseline condition in this wave, and we again calculated the mean differences between conditions two and three (table 6). The mean difference between conditions four and two is -3.83 percent. As in study 3, this result is not in the expected direction. Nevertheless, it is obvious that even in this condition the group of respondents did not change their responsiveness over time. As mentioned above, the difference between conditions one and four has changed over time. The difference has decreased and, due to this, the difference between conditions four and two is further reduced. However, the mean difference between conditions four and three has increased, with a value of 27.8 percent. This depends, on the one hand, on the increased mean difference between condition one and three and, on the other hand, on the decreased mean difference between conditions one and four. Therefore, one may assume that the agreement to the Islamophobia statement has changed from time one to time two.

Intraindividual stability

The next step to ensure the intraindividual stability of the instrument was to calculate the test-retest reliability of the four different conditions of the list experiment. The following table presents the two different perspectives (baseline with four and with five statements) plus the correlation coefficient between time one and time two (see table 7).

[Table 7 about here.]

The baseline condition with four statements indicates a high test-retest correlation of .696. This finding indicates that the respondents' provided quite stable answers in this condition from time one to time two. In contrast, the baseline condition with five statements leads to significantly more unexplained variance with r = .344, compared to condition one. From a methodological perspective this result is very interesting. It shows that adding a fifth nonsensitive statement significantly decreases the probability of respondents giving the same answer twice. One possible reason for this might be the larger number of statements.

Furthermore, the test-retest coefficient of the item condition "Muslims have too much influence in the world" with r = .456 is moderately stable. Basically, this finding means that the respondents did not substantially change their responses from time one to time two. However, the coefficient of the Jewish condition is quite small (r = .272). The respondents vary greatly in their responses over time. In this condition, the stability is especially low.

It is interesting that by comparing the means on the aggregate level, no changes were noticeable between wave 1 and wave 2. The two waves showed almost identical mean differences within the list experiment for condition two. The reason for the final result is that all respondents, as shown in the test-retest-reliability analyses, strongly revised their answers. This type of stability or instability is expressed by the test-retest correlation.

3. Discussion

Although the four studies presented here show inconsistent results, we do not ascribe them to a mode effect. In the literature, many studies exist which show that a social desirability response bias, especially for sensitive items, often appears in telephone or face-

to-face interviews but not in online surveys. For instance, Weisband and Kiesler (1996) show, by means of a meta-analysis, that self-administered online surveys produce less socially desirable distortion than interview-administered surveys (e.g., telephone). Also Tourangeau and Smith (1996) offer results which demonstrate that social desirability bias occurs more often within an interview situation.

In their studies of voter turnout reports that were carried out by using a comparison of the list experiment and direct self-report questioning, Holbrook and Krosnick (2010) found, that the list experiment could reduce the social desirability bias of voter turnout reports in the telephone survey. Indeed, no reduction could be detected between an online survey and direct questioning. The result of their analysis shows that social desirability or the social desirability pressure has no influence on voter turnout reports in online surveys.

However, we assume that such mode effects do not exist in our validation studies. Though, when comparing direct self-report questions and the results of the list experiment no differences appear in some cases. On the surface, study 1 showed promising results regarding anti-Semitism. Following the logic of the list experiment, socially desirable responsiveness can be supported in study 1. In contrast, in study 2 we found a significant difference ($p < .10$) in the opposite direction between the list experiment and the direct self-report question. This result was replicated (ns) in study 3. In both studies, the direct self-report question was more frequently responded to in the affirmative. Specifically, this finding means that respondents demonstrated more anti-Semitism in the obtrusive, direct question part compared to the unobtrusive list experiment. In addition, one has to take into account that the panel analyses, especially the test-retest reliability, showed that the stability of anti-Semitism is very low. Furthermore, study 3 (wave 1) demonstrated significant effects within the experiment in regard to Islamophobia and the very neutral item (“that one can easily make purchases at gas stations if necessary”). In these two conditions it is remarkable, as in study 1, that the

agreement to the direct self-report items is lower (ns) than to the unobtrusive measure. In the broadest sense, these results are in accordance with the logic of the list experiment. Nevertheless, they should not be considered as reliable because of the inconclusive panel data results. Furthermore, the list experiment is also not reliable, due to the controlling of the neutral (nonsensitive) statement, which should not induce a mean difference in the normal case. According to our results, the social desirability response set cannot easily and consistently be detected with the aid of the list experiment. The results of all three studies indicate that it makes no difference which mode of examination is used. In this way it becomes clear that in certain cases, a social desirability response bias might not be eliminated, and this neither in telephone surveys nor in online surveys.

A similar result to ours in study 3 was also found by Tsuchiya, Hirai and Ono (2007). They demonstrated that a social desirability response bias regarding shoplifting can appear in online surveys. Tsuchiya et al. showed that there are higher estimates in the list experiment (unmatched count technique) than in the direct question. However, in their study they also found that no difference appeared in the case of the nonsensitive item. This result is very contrary to our findings because we found that also the nonsensitive statement differs significantly. However, they explain the result by partial cultural differences. More specifically, they assume that the sensitivity or the opinion about shoplifting is clearly distinguished between the Japanese and the Western respondents (Americans). Nevertheless, the present findings of Tsuchiya and colleagues cannot provide answers for the inconsistent results in our study, especially in the field of anti-Semitism.

4. Conclusion

This paper has tested the extent to which the list experiment, including some modifications, yields valid, invariant, and stable results. On the basis of three studies and a total of five list experiments in two different modes, we have found that the list experiments

did not indicate consistent results. In addition, it was examined whether the mean increase is due to a higher number of items in the test condition. By means of a comparison between four and five neutral (nonsensitive) statements, a significant difference was found. Based on this finding, it can be deduced that the mean increased due to the higher number of items in the test condition. In other words, the higher the number of items, the higher the probability that more statements will be named. Furthermore, the five neutral (nonsensitive) statements were considered as the baseline condition and compared with the two test conditions in study 3. The mean differences in the two conditions are not significant but reveal an astonishing result, especially in the case of anti-Semitism. If one follows the logic of the list experiment, descriptively more respondents are angry about the item stating that one can easily make purchases at gas stations if necessary, compared to the item stating that Jews have too much influence in the world. Altogether, the list experiment shows very different results in the area of anti-Semitism. In study 1, a mean difference of almost 50 percent was demonstrated. In contrast, studies 2 and 3 indicate no significant and only minimal differences between baseline and test condition. Moreover, both studies show that the affirmation of the items in the direct question is only a few percentage points higher than in the indirect one (study 2, z = p < .10). This could be due to an underreporting effect. Tsuchiya and Hirai (2010) studied the causes which lead to such an underreporting effect. Their goal was to develop an appropriate list experiment that comes close to that of direct questioning. They argued that the instruction and implementation of the experiment does not fulfill the simple and common cognitive demands of the direct questioning and, thereby, underreporting effects arise. The reason for this phenomenon is that in the list experiment the respondents are only asked how many items they agree with and not, additionally, how many items they disagree with.[22] In such a format there is a lack of decision between “apply” and “does not apply” which appears in the

[22] Tsuchiya and Hirai (2010) use “applies” and “does not apply” in this case. Perhaps it could be possible to use other response formats.

case of direct questioning. In this way they are unable “to mentally consider both options equally” (Tsuchiya and Hirai 2010, p. 140). To confirm their hypothesis they compared several list experiments with different instructions with each other and with direct questions.[23] The result was that the list experiment[24], which included both options of “applies” and “does not apply” as instruction, was able to reduce the underreporting effect. However, it remains unclear how this technique acts with respect to sensitive items. Hence, regarding our studies, one could argue that the underreporting effect might have led to the inconsistent results concerning anti-Semitism. Furthermore, it is questionable whether a direct comparison between list experiment and direct questions seems useful in this context. The list experiment questions are about the feeling of anger and the direct question is concerned with the personal opinion or the agreement of the item. These are two different constructs which possibly allow no comparison and cause cognitive distortion. Moreover, it is not clear whether the respondents are angry about the "fact" that Jews have too much influence in the world or about the statement itself. In other words, why is someone able to create such a statement? This statement is disrespectful, and it is not right to think about Jews in that way. This possibility might have also led to distortions and inconsistent results. Furthermore, the panel analyses from study 3 yielded a low intraindividual stability of the instrument. The test-retest correlation of Islamophobia was .456 and significant whereas the test-retest correlation of anti-Semitism (.272) was not even significant.

Finally, we can summarize that the list experiment does not seem to be a robust instrument to control social desirability response bias regarding prejudices. Our three studies, with different outcomes, illustrate that this instrument shows considerable inconsistency. In nearly every study we discovered different results. From a social desirability perspective, the question remains of whether it is the sensitive item that causes the cognitive distortions. The

[23] They examined only nonsensitive items.
[24] In the paper it is called “elaborate item count technique.”

list experiment seems to result in a new black box for survey research. Future research must investigate the cognitive processes induced by the list experiment, especially in regard to sensitive issues, and to clarify under which conditions it works. This could be best done by combining cognitive interviewing with list experiments. Furthermore, experimental approaches could deepen our understanding of the list experiment. Assumed moderators like the number of statements, introduction wording, or the distorting effect of the sensitive item could be manipulated systematically.

References

Ahart, A. M., & Sackett, P. R. (2004). A New Method of Examining Relationships between Individual Difference Measures and Sensitive Behavior Criteria: Evaluating the Unmatched Count Technique. *Organizational Research Methods, 7*(1), 101–114.

Auspurg, K., Jann, B., Krumpal, I., & von Hermanni, H. (2012). Randomized-Response-Technik: Hope or Hype? Eine Meta-Analyse unter Berücksichtigung von Publication-Bias [Randomized-Response-Technique: Hope or Hype? A Meta-Analysis in Consideration of Publication Bias*]. Paper presented at the First Mini-Conference of the Center of Quantitative Methods of the University of Leipzig. Asking Sensitive Questions: Theory and Data Collection Methods.*

Blair, G., & Imai, K. (2012). Statistical Analysis of List Experiments. *Political Analysis, 20*(1), 47–77.

Chaudhuri, A., & Christofides, T. C. (2007). Item Count Technique in Estimating the Proportion of People with a Sensitive Feature. *Journal of Statistical Planning and Inference, 137*(2), 589–593.

Coutts, E., & Jann, B. (2011). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods & Research, 40*(1), 169–193.

Crowne, D. P., & Marlowe, D. (1960). A New Scale of Social Desirability Independent of Psychopathology. *Journal of Consulting Psychology, 24*, 349–354.

Dalton, D. R., Daily, C. M., & Wimbush, J. C. (1997). Collecting "Sensitive" Data in Business Ethics Research: A Case for the Unmatched Count Technique (UCT). *Journal of Business Ethics, 16*, 1049–1057.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the Unmatched Count Technique (UCT) to Estimate Base Rates for Sensitive Behavior. *Personnel Psychology*, *47*(4), 817–828.

Davidov, E., Thörner, S., Schmidt, P., Gosen, S., & Wolf, C. (2011). Level and Change of Group-Focused Enmity in Germany: Unconditional and Conditional Latent Growth Curve Models with Four Panel Waves. *AStA Advances in Statistical Analysis, 95*(4), 481–500.

Dovidio, J. F., & Gaertner, S. L. (1986). Prejudice, Discrimination, and Racism: Historical Trends and Contemporary Approaches. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, Discrimantion and Racism* (pp. 1–34). San Diego, CA: Academic Press.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991. The Item Count Technique as a Method of Indirect Questioning: A Review of its Development and a Case Study Application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Wiley Series in Probability and Statistics* (pp. 185–210). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Gilens, M., Sniderman, P. M., & Kuklinski, J. H. (1998). Affirmative Action and the Politics of Realignment. *British Journal of Political Science, 28*(1), 159–183.

Glynn, A. (2013). What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment. *Public Opinion Quarterly, 77*(S1), 159–172.

Heitmeyer, W. (2002). Deutsche Zustände, Folge 1.[German States. Vol. 1]. Frankfurt am Main: Suhrkamp.

Himmelfarb, S., & Lickteig, C. (1982). Social Desirability and the Randomized Response Technique. *Journal of Personality and Social Psychology, 43*(4), 710–7.

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring Voter Turnout by Using the Randomized Response Technique: Evidence Calling Into Question the Method's Validity. *Public Opinion Quarterly, 74*(2), 328–343.

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique. *Public Opinion Quarterly, 74*(1), 37–67.

Huddy, L., & Feldman, S. (2009). On Assessing the Political Effects of Racial Prejudice. *Annual Review of Political Science, 12*(1), 423–447.

Imai, K. (2011). Multivariate Regression Analysis for the Item Count Technique. *Journal of the American Statistical Association, 106*(494), 407–416.

Janus, A. L. (2010). The Influence of Social Desirability Pressures on Expressed Immigration Attitudes. *Social Science Quarterly, 91*(4), 928–946.

Jones, E. E., & Sigall, H. (1971). The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude. *Psychological Bulletin, 76*, 349–364.

Kane, J. G., Craig, S. C., & Wald, K. D. (2004). Religion and Presidential Politics in Florida: A List Experiment. *Social Science Quarterly, 85*(2), 281–293.

Karlan, D., & Zinman, J. (2011). List Randomization for Sensitive Behavior: An Application for Measuring Use of Loan Proceeds. *Working Paper.* Retrieved October 10, 2012, from http://karlan.yale.edu/p/JDE-ListRandomization.pdf.

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial Attitudes and the "New South". *The Journal of Politics, 59*(2), 323–349.

Kuklinski, J. H., Sniderman, P. M., Knight, K., Piazza, T., Tetlock, P. E., Lawrence, G. R., & Mellers, B. (1997). Racial Prejudice and Attitudes toward Affirmative Action. *American Journal of Political Science, 41*(2), 402–419.

Krumpal, I. (2011). Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality and Quantity. (Online First)* doi: 10.1007/s11135-011-9640-9.

LaBrie, J. W., & Earleywine, M. (2000). Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-Count Technique. *Journal of Sex Research, 37*(4), 321–326.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., & van der Heijden, P. G. M. (2005). Meta-Analysis of Randomized Response Research: 35 Years of Validation. *Sociological Methods & Research, 33*(3), 319–348.

Miller, J. D. (1984). *A New Survey Technique for Studying Deviant Behavior.* George Washington University.

Nickel, B., Berger, M., Schmidt, P., & Plies, K. (1995). Qualitative Sampling in a Multi-Method Survey. *Quality & Quantity, 29*(3), 223–240.

Oberski, D., Weber, W., & Révilla, M. (2012). The Effect of Individual Characteristics on Reports of Socially Desirable Attitudes towards Immigration. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, Theories, and Empirical Applications in the Social Sciences. Festschrift for Peter Schmidt.* Wiesbaden: Springer VS.

Papastefanou, G., & Wiedenbeck, M. (1998). Singuläre und multiple Imputation fehlender Einkommenswerte: ein empirischer Vergleich. [Singular and Multiple Imputation of Missing Income Data: An Empirical Comparison]. *ZUMA-Nachrichten, 22*(43), 73–89.

Paulhus, D. L. (1984). Two-Component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology, 46*(3), 598–609.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and Denial in Socially Desirable Responding. *Journal of Personality and Social Psychology, 60*(2), 307–317.

Raghavarao, D., & Federer, W. T. (1979). Block Total Response as an Alternative to the Randomized Response Method in Surveys. *Journal of the Royal Statistical Society. Series B (Methodological), 41*(1), 40–45.

Salzborn, S. (2010). The Politics of Antisemitism. *Journal for the Study of Antisemitism, 2*(1), 89–114.

Schuessler, K. F. (1982). *Measuring Social Life Feelings* (1st ed.). San Francisco: Jossey-Bass.

Schwarz, N., & Sudman, S. (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (1st ed.). San Francisco: Jossey-Bass Publishers.

Sniderman, P. M., & Grob, D. B. (1996). Innovations in Experimental Design in Attitude Surveys. *Annual Review of Sociology, 22*(1), 377–399.

Streb, M. J., Burrell, B., Frederick, B., & Genovese, M. A. (2008). Social Desirability Effects and Support for a Female American President. *Public Opinion Quarterly, 72*(1), 76–89.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response.* Cambridge, U.K.: New York: Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly, 60*(2), 275.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin, 133*(5), 859–883.

Tsuchiya, T., Hirai, Y., & Ono, S. (2007). A Study of the Properties of the Item Count Technique. *Public Opinion Quarterly, 71*(2), 253–272.

Tsuchiya, T., & Hirai, Y. (2010). Elaborate Item Count Questioning: Why Do People Underreport in Item Count Responses? *Survey Research Methods, 4*(3), 139–149.

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60*, 63–69.

Weisband, S. & Kiesler, S. (1996). Self Disclosure on Computer Forms. *Paper presented at the SIGCHI conference on Human factors in computing systems common ground – CHI´96,* pp. 3–10. *ACM Press*.

Zick, A., Wolf, C., Küpper, B., Davidov, E., Schmidt, P., & Heitmeyer, W. (2008). The Syndrome of Group-Focused Enmity: The Interrelation of Prejudices Tested with Multiple Cross-Sectional and Panel Data. *Journal of Social Issues, 64*(2), 363–383.

Zigerell, L. J. (2011). You Wouldn't Like Me When I'm Angry: List Experiment Misreporting. *Social Science Quarterly, 92*(2), 552–562.

Zimmerman, R. S., & Langer, L. M. (1995). Improving Estimates of Prevalence Rates of Sensitive Behaviors: The Randomized Lists Technique and Consideration of Self-Reported Honesty. *Journal of Sex Research, 32*(2), 107–117.

Acknowledgement:

The authors thank Uli Wagner and the members of the graduate school `Group-Focused Enmity` for their helpful comments on an earlier draft of this article. This work was supported by the Volkswagen Foundation, Freudenberg Foundation and Möllgaard Foundation (financial support to the project on `Group-Focused Enmity`). The research of Peter Schmidt was supported by the basic research program of the State Research University Higher School of Economics (HSE) in Moscow. The authors would also like to thank Lisa Trierweiler for the English proof of the manuscript.

Appendix A: Methodological Details

Study 1 and Study 2

Interviewing for study 1 was done between April 17, 2009, and June 29, 2009. Data for study 2 were collected between June 5, 2010, and August 14, 2010. The basic population consists of all Germans that are over the age of 16. For random digit dialing the sample was drawn from a database that contains approximately 36 million generated landline numbers. These numbers based on 360.000 blocks of at least one listed number per hundred blocks. There was no previous screening or cleaning method used. The respondents were selected within the households by the Kish-grid method.

Based on the Standard Definitions the refusal rates are:

Study 1 (2009)

RRF1 0,164

RRF2 0,276

RRF3 0,473

Study 2 (2010)

RRF1 0,184

RRF2 0,212

RRF3 0,060

Table 1. Overview of Different Studies.

Study 1 conducted in 2009		**Mode**
Baseline Condition	Test Condition	CATI
3	3+Jews	N = 229
Study 2 conducted in 2010		
Baseline Condition	Test Condition	CATI
4	4+Jews	N = 445
Study 3 (wave 1) conducted in 2011		
Baseline Condition	Test Condition	ONLINE
4	4+Jews	N = 1,569
	4+Muslim	
	4+gasoline station*	
Study 3 (wave 2) conducted in 2011		
Baseline Condition	Test Condition	ONLINE
4	4+Jews	N = 194
	4+Muslim	
	4+gasoline station*	

*nonsensitive item

Table 2. Estimated Mean Level of Anger for Baseline and Test Condition.

Baseline condition (3 items)	Test condition (4 items) "…that Jews have too much influence in the world."	Difference between baseline & test condition Percent "angry" (95% confidence interval)
1.39 (.105)[a] (N = 108)	1.88 (.094) (N = 121)	48.62% (14.1)*** (20.73 to 76.44%)

[a]Standard error of the estimate
*p < 0.05, p** < 0.01, ***p < 0.001

Table 3. Difference of Proportion Test: Comparison of List Experiment and Direct Question.

	Study 1	Study 2	Study3/ Anti-Semitism	Study3/ Islamophobia
Difference between baseline & test condition (estimation of anger)	48.62% (14.1)[a]***	9.48% (9.9)	8.40% (7.2)	33.69% (7.1)***
Sample size	229	445	755	780
Direct question (agreement)	11.7%	24.1%	13.7%	26.8%
Sample size	94	195	323	385
Z-score[b]	**2.52****	**1.41***	**1.25**	**0.92**

[a]Standard error of difference
[b]Z-score indicates the significance difference of the proportion from the list experiment and the direct question
*p < 0.10, **p < 0.01, ***p < 0.001

Table 4. Estimated Mean Level of Anger for Baseline and Test Condition.

Baseline condition (4 items)	Test condition (5 items) "…that Jews have too much influence in the world."	Difference between baseline & test condition Percent "angry" (95% confidence interval)
1.96 (.075)[a] (N = 226)	2.06 (.065) (N = 219)	9.48% (9.9) (-10.00 to 28.95%)

[a]Standard error of the estimate
*p < 0.05, **p < 0.01, ***p < 0.001

Table 5. Estimated Mean Level for Four Different Conditions (Waves 1 and 2).

Wave 1	**Condition 1**	**Condition 2**	**Condition 3**	**Condition 4**
	4 statements (N = 376)	4 statements + Jews (N = 379)	4 statements + Muslims (N = 404)	4 statements + gasoline station (N = 410)
	1.71 (.045)[a]	1.79 (.054)	2.05 (.052)	1.93 (.048)
Wave 2	**Condition 1**	**Condition 2**	**Condition 3**	**Condition 4**
	4 statements (N = 50)	4 statements + Jews (N = 44)	4 statements + Muslims (N = 49)	4 statements + gas station (N = 51)
	1.78 (.141)[a]	1.86 (.144)	2.18 (.142)	1.90 (.129)

[a] Standard error of the estimate

Table 6. Estimated Mean Difference of Four Conditions (Waves 1 and 2).

Wave 1 Condition 1	**Conditions 2, 3, and 4**	**Difference between conditions 1 and 2, 3, 4 Percent "angry" (95% confidence interval)**
4 statements	4 statements + Jews	8.40% (7.2)[a] (-10.10 to 26.92%)
	4 statements + Muslims	33.69% (7.1)*** (15.46 to 51.91%)
	4 statements + gas station	22.40% (7.1)** (4.24 to 40.56%)
Condition 4	**Conditions 2 and 3**	**Difference between conditions 4 and 2, 3 Percent "angry"**
4 statements + gas station	4 statements + Jews	-14.00% (7.0) (-32.11 to 4.29%)
	4 statements + Muslims	11.28% (6.9) (-6.54 to 29.11%)
Wave 2 Condition 1	**Conditions 2, 3, and 4**	**Difference between conditions 1 and 2, 3, 4 Percent "angry" (95% confidence interval)**
4 statements	4 statements + Jews	8.81% (20.1)[a] (-43.23 to 60.86%)
	4 statements + Muslims	40.45% (19.34) (-4.39 to 85.29%)
	4 statements + gas station	12.65% (19.34) (-31.98 to 57.27%)
Condition 4	**Conditions 2 and 3**	**Difference between conditions 4 and 2, 3**
4 statements + gas station	4 statements + Jews	-3.83% (19.89) (-49.73 to 42.06%)
	4 statements + Muslims	27.80% (19.24) (-16.59 to 72.20%)

[a] Standard error of the difference
*p < 0.05, **p < 0.01, p*** < 0.001

Table 7. Test-Retest Reliability (Waves 1 and 2).

Methodological	**Correlation coefficient Between times 1 and 2**	**N**
Baseline 4 statements	.696**	50
Baseline 5 statements	.344*	51
Substantial	**Correlation coefficient Between times 1 and 2**	**N**
Muslim condition	.456**	49
Jews condition	.272	44

*p < 0.05, **p < 0.01, ***p < 0.001

Manuscript #2:

Cognitive Distortions in the List Experiment:

A Mixed Method Approach

Stefanie Gosen

Philipps-University Marburg, Germany

Submission date: January 20, 2014

Abstract:

The list experiment is an indirect survey method, which should inhibit socially desirable response bias by providing a higher level of anonymity and privacy. Especially for sensitive issues such as own drug use, prejudice, etc., the list experiment should provide higher estimates of the proportion of people who have engaged in the sensitive item compared to direct self-reports. In the literature, however, list experiments do not always follow this hypothesis, and direct self-reports yield higher estimates. I performed one qualitative and two experimental studies to examine the factors explaining these inconsistent results. The cognitive interviews showed that a few respondents noticed the sensitive item but the nonsensitive items were easy to understand and it was easier to provide an answer for those items. In total, the procedure of the list experiment was understood by almost all respondents. The first experimental online study showed that the response behavior changed when a sensitive item was included. Moreover, the approval rate of the nonsensitive items increased significantly when a sensitive item was included. The result reached via the indication of the number of *yes* answers by the respondents is significantly higher than direct self-reports, regardless of whether a sensitive item was included or not. Three factors are pointed out which might possibly provide answers to the inconsistent results of the list experiment: (1) the different number of items in the different conditions; (2) the item difficulty is distorted by a sensitive item, (3) the procedure to count the number of items is biased.

Keywords: indirect questioning, list experiment, social desirability bias, sensitive questions, item difficulty, mixed method

1 Introduction

Social desirability response biases are a common problem in surveys. The problem increases when it comes to particularly sensitive issues, for instance, drug use, prejudice against minorities, sexual behavior, abortion, etc. (Droitcour et al. 1991; Jones and Forrest 1992; Krumpal 2012; Krysan and Couper 2003; Tourangeau and Yan 2007). Social desirability biases often appear in direct self-reports in which the respondents are asked to give honest answers. This situation mostly leads to impression management in order to present oneself in a positive light to the interviewer (Paulhus 1984) and is linked with adaption of social norms and the degree of privacy in an interview situation (Tourangeau et al. 2000, p. 257). Therefore, respondents adapt a socially desirable opinion, which results in systematic distortions (Krumpal 2013; Tourangeau and Yan 2007). To counteract this problem, more indirect and unobtrusive survey methods are used and developed, e.g., implicit association test, bogus pipeline, nominative technique etc. (Greenwald et al. 1998; Roese and Jamieson 1993; Sirken 1970). These techniques guarantee anonymity and privacy or placed the respondent in a situation in which the researcher receives more honest and valid self-reports with regard to sensitive issues (Chaudhuri and Christofides 2007; Paulhus 1984; Warner 1965).

A commonly used indirect method is the randomized response technique (Warner, 1965). By providing the opportunity for respondents to randomize their answers, this technique guarantees them anonymity (Krumpal 2013). In this paper I am going to focus on a method that is very similar to the randomized response technique, the list experiment (Kuklinski et al. 1997a; Kuklinski et al., 1997b; Sniderman and Grob 1996). It was first presented by Miller (1984) as the *unmatched count technique*. It is also known as the *item count technique* (Dalton, Wimbush, & Daily, 1994), the *randomized list technique*

(Zimmerman and Langer 1995), and the *list randomization technique* (Karlan and Zinman 2012).

The aim of the list experiment is to receive a correct estimation of the agreement to a sensitive item by guaranteeing anonymity. In order to show socially desirable response bias, this estimate (calculated on the aggregate level) is compared to the direct self-report questions. The list experiment proceeds as follows: The respondents are randomly split into two groups. While the baseline condition receives three items that only contain nonsensitive content, the test condition receives the same nonsensitive items plus one additional sensitive item. In both groups, the respondents are asked to indicate the number of statements that make them angry. They should only report the number of statements that elicited anger, and it is important that they only report the number of and not the specific statements. Otherwise, anonymity could no longer be guaranteed. Here is an example of the list experiment performed by Kuklinski et al. (1997b, p. 405) in the field of racism:

> Now I'm going to read you three/*four* things that sometimes make people angry or upset. After I read all three/*four*, just tell me how many of them upset you. I don't want to know which ones, just how many.
>
> 4. the federal government increases the tax on gasoline;
> 5. professional athletes getting million-dollar salaries;
> 6. large corporations polluting the environment.
> 7. *a black family moving in next door.*

Additionally, respondents in the test condition received the following sensitive item: *a black family moving in next door.* Thus, the list experiment is able to generate an estimate of the number of respondents who got angry about the sensitive item (Dalton et al. 1994; Kuklinski et al. 1997b). This estimate is only possible to calculate in the aggregate level. Thereafter, the mean level of the baseline condition is subtracted from the mean level of the test condition. In order to receive a percentage value, the mean difference of the two conditions were multiplied by 100. The logic of the list experiment implies that every

increase in the mean of the test condition is considered to show agreement to the sensitive item. For instance, if the baseline condition had a mean of 2.5, and the test condition mean was 3.0, the difference of both conditions would be 0.50 (*100 = 50 percent). In other words, 50 percent of the respondents are assumed to have become angry about the sensitive statement and would have expressed their prejudices against blacks (Streb et al. 2008, p. 81).

2 State of research and the inconsistency of the list experiment

There are numerous studies that show that the list experiment uncovered more socially undesirable behavior than direct self-reports (e.g., Blair and Imai 2012; Coutts and Jann 2011; Dalton et al. 1994; Kane et al. 2004; Kuklinski et al. 1997b; Miller 1984; Rayburn et al. 2003; Streb et al. 2008; Wimbush and Dalton 1997). Nonetheless, among the many positive findings, there are studies that exhibit insufficient outcomes in the area of social desirability response bias. This means that the list experiment was not able to indicate higher estimates compared to those achieved by direct self-report survey questions. For instance, in a study of HIV risk behavior, Droitcour et al. (1991) found that direct self-reports received higher estimates than the list experiment. On the issue of drug use, Biemer et al. (2005) also could not find higher estimates using the list experiment compared to direct-self reports. Furthermore, the study of counterproductive behavior by Ahart and Sackett (2004) was unable to uncover evidence of higher estimates of the list experiment compared to the direct self-reports. A further study, which reflected the inconsistent results of the list experiment, was presented by Holbrook and Krosnick (2010). Here, the list experiment yielded higher estimates of voter turnouts in telephone surveys. In addition, the list experiment was conducted in three online surveys; when surveyed in this mode, the direct self-reports provided higher estimates than the list experiment. In contrast to that study, in their online survey regarding shoplifting, Tsuchiya, Hirai, and Ono (2007) found that the list experiment received higher estimates than direct reports.

In the field of prejudice research, Gosen et al. (in prep.) provide further inconsistent results. They conducted and compared three studies implementing a total of five list experiments and two different types of survey modes. The aim of the entire analysis was to test the external validity, such as the invariance of the results across the different studies in the field of prejudice.

The first study exhibited promising results regarding anti-Semitism. The authors found a significant difference between the test and baseline condition of the list experiment. In addition, the list experiment received higher estimates compared to the direct self-reports. In contrast, a modified repetition of the first study showed no significant difference in the list experiment. Furthermore, the direct self-report item yielded higher estimates than the list experiment. In the second study, the list experiment was unable to uncover a socially desirable response bias. On the basis of these two inconsistent outcomes, a further online study was conducted. This study compared three different list experiments with each other. The first list experiment, covering anti-Semitism, demonstrated nearly similar results as the second study. Here, no difference was found between test condition and baseline condition in the list experiment regarding anti-Semitism and the direct self-reports had higher estimates than the list experiment. To check for validity, another prejudice item was included as a sensitive statement in the second list experiment. This item covered the field of Islamophobia. In this case, there was a significant difference in the expected direction. The list experiment yielded higher estimates than the direct self-reports. A central question in this survey was if the change or increase in the mean was caused by the higher number of items in the test condition. In this respect, in the third list experiment, another nonsensitive statement was added to the four basic nonsensitive items in the test condition. This item was conceived to be very neutral and provide less agreement. The item was as follows: “That one can easily make purchases in gasoline stations if necessary”. This third list experiment yielded a

significant difference in the expected direction. In this case, if one follows the logic of the list experiment, no difference should have occurred, and the test condition should not have had a higher mean level as the baseline condition. In sum, the number of the items in the different conditions influenced the mean in the test condition.

According to the findings of the studies considered here, no consistent results have been achieved. Considering this varying background, the questions arise of how the respondents generate their responses and how the cognitive process functions. I assume that three moderating factors might explain the inconsistency.

One factor, which was tested in the abovementioned studies from Gosen et al. (in prep.), is that the mean increased in the test condition due to a higher number of items. In other words, the higher the number of items, the higher the probability that the respondents named more statements. The other two factors that might be responsible for the inconsistent results found up to now should be answered in this paper.

2. The first research question (second factor) is whether a distortion in the item difficulty exists (Gilovich et al. 2002)? In other words, do the respondents notice the sensitive item in the list experiment and, if so, do they shift their response patterns to the nonsensitive questions? In this regard, the nonsensitive items are focused on by respondents because these items are easier to answer in relation to the sensitive items. The sensitive item could change the valence of the nonsensitive items to such an extent that, for instance, the approval or rejection rates will increase or decrease in a list experiment, thus distorting the mean level in the test condition in the list experiment.

3. The second research question (third factor) investigates whether the procedure used to count the number of items to which the respondents reply in the affirmative might be distorted. It could be possible that the respondents are not able to keep the items in mind or

that they do not really count the number of statements they are angered by but rather give a general (not thought about) answer regarding the number of items. Therefore, the list experiment's mean level in the test and baseline condition might be affected by upward or downward bias.

Hence, this paper seeks to provide an answer to the general research question of whether the inconsistency of the list experiment can be proved by manipulating the two last moderating factors.

3 This Paper

After extensive research including a specific analysis of the validity of the list experiment (Gosen et al. in prep.), it is often not evident why the list experiment has provided different results. In this paper, extensive focus has been placed on *qualitative and experimental designs* for several reasons: to deepen the functionality of the list experiment, to ascertain the respondent's understanding of the list experiment, and to test the effect of the sensitive item. On the basis of three studies, ambiguous areas of the list experiment will be analyzed more thoroughly. Study 1 deals with the general understanding on the qualitative level. Via cognitive interviews (Willis 2005), the implementation and understanding of the list experiment is distinctly captured. The focus will also be on the sensitive item in the list experiment. This particular item might be responsible for cognitive distortions on the part of the respondents because of the enhanced item difficulty. Study 2 is an experimental design and was conducted in the form of an online survey. The main focus of this study is on capture the item difficulty and on whether the induction of the sensitive item affects or distorts the approval rate of the nonsensitive items. Furthermore, the extent to which the indication of the number of *yes* answers (which will be referred to hereafter as the 'individual aggregate value') is biased by varying the position of the sensitive item. The study also examines whether distortions occur between the individual aggregate value and the mean level of the

sum of *yes* answers in the direct self-report items (which will be referred to hereafter as the 'sum of direct *yes* answers'). The actual answer to the number of items that anger the respondents may potentially be another factor that causes distortions, and thereby questions the validity of the list experiment. The last study is a replication of Study 2. Study 3 should replicate the results of Study 2 by controlling for repeated measurement problems (for an overview of the studies see Table 1). The added value of this paper is the mixed method approach combining qualitative (cognitive interviews) and experimental methods, giving new insights into the response process elicited by the list experiment.

-Table 1 about here-

4 Study 1: Cognitive Interviews

On the basis of the former studies and the inconsistent results of the list experiment, I decided to conduct cognitive interviews in order to better understand the cognitive processes of the respondents in response to the items. Furthermore, I developed a particular catalog of questions for the list experiment. By means of cognitive interviews, I examined the following questions:

1. Do the respondents notice the sensitive item in the list experiment? Are they able to distinguish clearly between sensitive and nonsensitive items?
2. Do the respondents express anger about the content of the sensitive question or about the fact that the interviewer asked such a delicate question?
3. Is the list experiment instruction difficult to understand when it is read aloud to the respondents?

Method

Participants and procedure

I conducted a series of seven interviews with respondents who answered an advertisement in an online magazine. The respondents were four women and three men, and their mean age was 41.42 years (SD = 15.6). Regarding educational level, participants showed an acceptable demographic range: Specifically, the represented levels of education included lower secondary education (1 respondent), secondary education (4 respondents), upper secondary education (1 respondent), and college graduate (1 respondent).

The list experiment was integrated in an extensive questionnaire consisting of direct self-report items of the syndrome of Group Focused Enmity (GFE) (Heitmeyer 2002; Zick et al. 2008) and demographic items. The list experiment was carried out alternately at the beginning or at the end of the questionnaire. This scheme was adopted as it should help to reduce any sequence effect or show whether the answer of the respondents varied with respect to the different positions. In addition, I conducted two different conditions of the list experiment. In the first condition, the sensitive item was placed at the top of the list, and in the second condition, it was placed at the bottom. In total, four different questionnaires were tested (see Table 2).

The question wording of the list experiment was as follows:

> Now I will refer to some topics people occasionally express anger about. Could you please tell me how many of the following five statements you have also been angry about? Please count, using your fingers, how many have angered you.
>
> 1. The way gas prices keep going up;
> 2. That professional athletes are getting million-dollar salaries;
> 3. That the German railroad has so many delays [1];
> 4. That seat belts are required to be used when driving[2];
> 5. That Jews have too much influence in the world. (Alternatively: Muslims

[1] The items were used in the studies of Gosen et al. (in prep.).
[2] The original item is taken from Streb et al. (2008).

"Have you ever been angry about five, four, three, two, one, or none of the statements mentioned?"

-Table 2 about here-

Each questionnaire was read out by the interviewer (face-to-face), and the respondents had to give their answers verbally. I assumed this way to be one of the best to test how well the respondents understood the list experiment and to directly observe if they had any problems with the instructions. However, it turned out to be more difficult for respondents when they had to listen to the interviewer than when they could read the question by themselves. This situation imposed more of a cognitive burden than self-administered questionnaires (Tourangeau et al. 2000; Chaiken and Eagly 1976) where they read the items themselves. I supposed that if the respondents were capable to understand the list experiment in a read-aloud procedure they would also understand it by reading it by themselves, e.g., in paper-pencil or online surveys.

In the interviews I used two established cognitive interviewing techniques – the probing technique (Willis et al. 1999) and the think-aloud method (Willis 2005). In the first technique, probe questions are asked by the interviewer after respondents answered the question, i.e., question of understanding or answer process in general (Willis et al. 1999). When using the think-aloud method, the respondents should explain their thoughts while they answering the question (Willis 2005). Especially the latter was supposed to provide more information for the list experiment. The respondents were asked to describe their train of thought during the experiment. This should bring insight to cognitive processes that contribute to the respondent's answers. Unfortunately, this method could not be implemented at the data collection stage. The main problem was that the respondents did not understand

the procedure or were unable to articulate their thought processes. Therefore, it was necessary to alternatively use the probing technique. However, to test my assumptions, I developed specific questions for the list experiment. The questions were as follows[3]: "Was it difficult for you to answer the aforementioned question (list experiment)?", "How difficult was it?" (five-point scale with a range from very easy to very difficult), "Why was it that easy / that difficult?", "Do you have a presumption of what the researcher wanted to find out with such a question?", "Did you find the individual statements strange?", "Have you wondered about the fact that the question was asked for so many issues?" In addition, some spontaneous questions were added by the interviewer.

Results

Based on the results of the cognitive interviews, the first assumption about whether the respondents noticed the sensitive item and distinguished clearly between the nonsensitive and the sensitive item can be partly confirmed. In three interviews, the sensitive item was perceived, but not as a negative question. Only one respondent mentioned that all five statements of the list experiment did not fit together thematically. Furthermore, in these three interviews, the respondents could distinguish clearly between the sensitive item and nonsensitive items. However, in five interviews, respondents felt more comfortable answering the nonsensitive item of the list experiment than the sensitive direct self-report questions. This was due to the fact that the nonsensitive items included more personal issues which were easy to answer because of the personal experience of the respondents or experiences of family members or friends. Furthermore, findings strongly corroborated the second assumption of whether the respondents express anger about the content of the statement and not about the fact that the interviewer asked such a sensitive issue. Every respondent who got angry about the sensitive statement expressed anger on the content level.

[3] Only the specific items are shown in this paper. The other questions were general probing questions.

Moreover, the results for the third assumption indicated that the respondents had no problems understanding the list experiment and its task. In total, I found six respondents who had no problems understanding the introduction or the task of the list experiment. Only one out of seven respondents reported difficulties answering the list experiment. The formulation of the questions seemed to be obvious to the respondents. Another striking point was that no respondents questioned the purpose of the list experiment. Furthermore, only one person indicated understanding what the researcher wanted to find out. The respondent believed that the researcher wanted to know whether the people are lying or not. This description comes very close to the underlying intention of the list experiment. Overall, however, the majority of the respondents did not understand the underlying intent of the list experiment.

Discussion

The results indicate that, during the interviews, no distortions became apparent when answering the list experiment. In nearly every interview, there were no problems of understanding or any other serious conspicuous issues. Indeed, these interviews were unable to reveal an unambiguous bias by the sensitive item because only three respondents were able to distinguish between the sensitive and nonsensitive items. The other four respondents were not consciously aware of the sensitive item. Nevertheless, with the interviews it was possible to show especially the respondents reactions, reason, and process of responses, and what they are thinking about the nonsensitive items. Thus, the results demonstrated that the nonsensitive items were easier to answer. In addition, this could be a first indication that the respondents shifted their response pattern to the nonsensitive items. In other words and according to above-mentioned research question, the item difficulty of the nonsensitive items can be changed when a sensitive item is included. It seems that the different valences are one of the key aspects, when it comes to distortions. I assume that the change of the item difficulty is not perceived consciously by the respondents, but rather expressed unconsciously in the

agreement to the nonsensitive items. In the interviews, the sensitive item attracted the attention of only three respondents. Nevertheless, the shifted item difficulty cannot be excluded in common questionnaires and list experiments. This assumption cannot be proven in the cognitive interviews, but is a very important factor for further research. To test the possible distortion of the nonsensitive question due to the sensitive question, it is necessary to use a quantitative experimental design. In Studies 2 and 3, the change of the response pattern will be tested (change of item difficulty). In addition, it will be also checked if the procedure to count the number of *yes* answers leads to possible distortions since this issue was not tested in the cognitive interview design. Finally, it should be mentioned that the inconsistent results mentioned in the introduction are not directly related to an incorrect or false understanding of the list experiment.

5 Studies 2 and 3: Experimental Design

Study 2

Study 2 provides a first test of the change of item difficulty and the difference between the individual aggregate value and the direct self-report items. In an online experiment, I manipulated the position of the sensitive item.

Method

Participants and procedure

Data were collected in March 2013. Participants were recruited via the mailing list via the mailing list of the University of Marburg. The sample size was N = 1,878 (50% had completed upper secondary education). Of the total sample, 1,018 were female. Participants' mean age was 29.96 years (SD = 10.35) and ranged from 14 to 70 years.

The respondents were allocated randomly into control and experimental condition.[4] In general, the questionnaire consisted of three different segments, hereafter referred to as blocks. One block consists of sensitive and nonsensitive items, which are presented as direct self-report questions with each question placed on a separate page (for an overview see Table 3). Every respondent in control and experimental condition passed each block. The first block included four nonsensitive items and one sensitive item. The response scale was dichotomous consisting of the two options yes and no (for item content see Table 4). The sensitive item in block 1 was as follows: "I am angry about Jews having too much influence in the world." (Item content of the nonsensitive items is shown in Table 4). After the five single item pages, participants were asked to count the number of items they agreed to. The original question in the questionnaire was as follows: "Please indicate how many of the previous questions you answered with *yes*." As in the cognitive process of the list experiment, they had to count the number of *yes* answers (and based on this number, the individual aggregate value will be formed). In block 2, the structure of the pages was the same as in block 1, but the content of the items was changed. The sensitive statement in this block was, "I am angry about Muslims having too much influence in the world." At the end of this block, respondents were not asked to indicate the number *yes* answers (individual aggregate value) to avoid learning effects. Furthermore, distractors were placed between the individual blocks. These were simple IQ test tasks that should direct the attention of the respondents in another direction. This was done to counteract possible distortions that might arise due to the repetition a certain type of question.

-Table 3 about here-

[4] The random assignment in terms of gender, age, and education was successful across the conditions.

In contrast to the first two blocks, the last block (block 3) contained five nonsensitive items. Again, different items than in block 1 and 2 were used. In block 3, the respondents had to indicate the number of *yes* answers.

-Table 4 about here-

In the control condition (N = 948), the sensitive items were presented after the nonsensitive items in blocks 1 and 2. In the control condition, the nonsensitive items could not be affected by the sensitive item because it was presented on the last page of each block, and thus, the possible influence on the approval or rejection rate of the respondents could be controlled for. In block 3, the positioning of the nonsensitive 'control' item, "I am angry that one can easily make purchases in gasoline stations if necessary" was changed. In the control condition, this particular item was placed after the other nonsensitive items, as the sensitive item in blocks 1 and 2.

In the experimental condition (N = 930), the sensitive item was included. This means it was placed prior to the nonsensitive items and was presented first. Accordingly, the approval or rejection of the nonsensitive items should be affected by the sensitive item on the first page. Also in block 3, the fifth nonsensitive item (see Table 4) was moved to the beginning of the block. This procedure guaranteed a certain comparability to blocks 1 and 2. At the very end of the questionnaire, sociodemographic variables were measured.

In this study, I address the following hypotheses:

Change of item difficulty:

1. The agreement to the nonsensitive items changes if a sensitive item is included.

   → Hypothesis is tested in blocks 1 and 3.

The procedure to count the number of *yes* answers (individual aggregate value) leads to further distortions:

2. The position of the sensitive item causes distortions in the individual aggregate value.

   → Hypothesis is tested in block 1.

3. The indication of the number of *yes* answers leads to distortions.

   → Hypothesis is tested in blocks 1 and 3.

Results: Item difficulty

H1: The agreement to the nonsensitive items changes if a sensitive item is included.

In the following analysis, I used only blocks 1 and 3. Block 2 was excluded from the analysis because it was used as a further distractor variable and did not contain the question to indicate the number of items with *yes* answers. To test the first hypothesis, a t-test between experimental and control condition in block 1 was applied. To compare exclusively the nonsensitive direct self-report items in the block of interest, only the sum of direct *yes* answers of four nonsensitive items without the sensitive item of each question were generated. I found a significant effect between the two conditions, $t(1866) = -1.98$, $p < .05$, showing that the experimental condition ($M = 1.77$, $SD = .89$) reached a higher mean level than the control condition ($M = 1.69$, $SD = .91$). To confirm the hypothesis and to test that the distortions in the experimental condition can be attributed to the sensitive item, a t-test was used in block 3 (only nonsensitive items). In this case, the mean levels were generated of the sum of direct *yes* answers of four nonsensitive items. The 'control' item, "I am angry that one can easily make purchases in gasoline stations if necessary", was not included in the analysis. No significant effect emerged between the two conditions, $t(1731) = -.831$, $p = .41$. The result clearly shows that when only nonsensitive items were presented, no biases

occurred between the control (M = 1.60, SD = 1.68) and the experimental condition (M = 1.67, SD = 1.63). The mean levels remained constant.

Results: Biased individual aggregate value

The hypothesis was:

H2: The position of the sensitive item causes distortions in the individual aggregate value.

To test the second hypothesis I conducted a 2 x 2 repeated measure ANOVA with the individual aggregate value and the mean level of the sum of direct *yes* answers as within-subject factors and experimental conditions 1 and 2 as between-subject factors.[5]

In experimental condition 1, the sensitive item was at the end of block 1 before the individual aggregate value. In experimental condition 2, the sensitive item was presented at the beginning of the block; a nonsensitive item was presented prior the individual aggregate value. The mean level of the sum of direct *yes* answers was established out of the five items (one sensitive + four nonsensitive) of each question. There was a significant main effect for the individual aggregate value and the mean level of the sum of direct *yes* answers, $F(1, 1374) = 250{,}83$, $p < .001$, partial $\eta^2 = 0.15$. Figure 1 displays that the means of the individual aggregate value in experimental condition 1 (M = 2.18, SD = 1.36) and experimental condition 2 (M = 2.24, SD = 1.36) are higher than the mean of the sum of direct *yes* answers (experimental condition 1: M = 1.80, SD = .94; experimental condition 2: M = 1.87, SD = .92). Deviating from the assumptions, the interaction effect between the two conditions was not significant, $F(1, 1374) = 0.016$, $p = .90$, $\eta^2 = 0.00$. The position of the sensitive item had no effect on the individual aggregate value.

[5] To avoid any problems of understanding with test and experimental condition of the test of the first hypothesis in this part of the analysis, the different conditions were named experimental condition 1 and experimental condition 2.

-Figure 1 about here-

The next step is to test:

H3: The indication of the number of yes answers leads to distortions.

To investigate the third hypothesis and to ascertain whether the individual aggregate value indicates general distortions, a 2 x 2 design and a repeated measure ANOVA was used. The within-subject factors were the individual aggregate value and the mean level of the sum of direct *yes* answers in block 1 and the individual aggregate value and the mean level of the sum of direct *yes* answers in block 3. Blocks 1 and 3 can be seen as between-subject factors.

In order to test the differences between individual aggregate value and the sum of the direct *yes* answers, the experimental (item prior) and control condition (item after) were summarized to only one condition. This especially makes sense since it was only the question about the difference of the individual aggregate value and of the sum of direct *yes* answers and not whether there were differences between the different conditions. Another reason to unify the experimental and control condition was that there were no differences in the positioning of the sensitive item. Hence, the mean level of the sum of direct *yes* answers was produced from all five items in each of the blocks. In the block with the sensitive item (block 1), the main effect was significant, $F(1, 1160) = 223.54$, $p = .000$, $\eta^2 = 0.16$. Figure 2 shows that the mean level of the individual aggregate value ($M = 2.19$, $SD = 1.36$) and the mean of the sum of direct *yes* answers ($M = 1.81$, $SD = .93$) differ significantly and presents an increased individual aggregate value. In contrast, the main effect in block 3 (only nonsensitive items) was nonsignificant, $F(1, 1160) = 0.002$, $p = .96$, $\eta^2 = 0.00$. The mean of

the sum of direct *yes* answers (M = 1.38, SD = 1.35) and the mean level of the individual aggregate value (M = 1.38, SD = 1.20) did not differ at all (see Figure 2).

-Figure 2 about here-

Discussion

The results of Study 2 concerning H1 indicate that the sensitive item biased the item difficulty of the nonsensitive items. If the sensitive item was presented prior to the nonsensitive items, the agreement to the nonsensitive items increased. This points out that the valences of the nonsensitive items were changed when a sensitive item was contained or added. The nonsensitive items seem to be easier to answer when a difficult sensitive item is included. In other words, the respondents set the anchor in the difficulty to the sensitive item and try to counteract the difficulty by agreeing more often to the "easier" nonsensitive items (Tversky and Kahneman 1974). This assumption can be confirmed with the analysis of block 3 in Study 2, in which there were only nonsensitive items in each condition. In this block, no difference was found. In both in the experimental and in the control condition the respondents agreed in the same way.

The findings of the second part of the Study cannot be confirmed (H2). I found no differences between the experimental conditions 1 and 2. The individual aggregate mean level increased in both conditions, in contrast to the mean level of the sum of direct *yes* answers. The respondents stated a higher individual aggregate number in experimental conditions 1 and 2. It did not make any difference if the sensitive item was presented at the beginning or at the end of the specific block. The positioning of the sensitive item showed no effect.

The present findings cannot corroborate the hypothesis of the distorted individual aggregate value. In this study, the first block with the sensitive item provided a higher mean level of the individual aggregate value as the mean level of the sum of the direct *yes* answers irrespective of whether the sensitive item was at the beginning or at the end of the block. This result might, at first sight, be plausible and in line with the functionality of the list experiment because the third block with only nonsensitive items exhibited no differences between the individual aggregate value and sum of direct *yes* answers. Especially these results seem to confirm the logic of the list experiment regarding the truthful answers in the aggregate format. The respondents had a feeling of anonymity by indicating only the number of their *yes* answers. For that reason it is logical that the aggregate value is higher than the sum value in the sensitive item condition.

In sum, Study 2 provides mixed findings and supports my predictions only in one part. However, the results are subject to the following limitations and problems. A substantial limitation is the problem of the repeated measurement, in which the respondents were presented "list experiment-like" blocks twice. In general and in traditional surveys, the common procedure of the list experiment is to survey the respondents as either a part of the test condition or the baseline condition and receive the list experiment only once. This process prevents learning or sequence effects. The results in block 3 might be biased because the respondents were more familiar with the kinds of questions and procedure in the last part of the questionnaire. As a consequence, the respondents could count their *yes* answers after the first block and no biased individual aggregate value may occur. At the beginning of the questionnaire, no instruction was given in which the respondents were informed to count the number of *yes* answers. After the first block, the participants might conclude from the first declaration of the individual aggregate value that further tasks will follow. Following this account, the respondents continued to count the number of *yes* answers, which distorted the

results. To counteract this problem a further study was conducted. This study (Study 3) was replicated with blocks 1 and 3 being switched in order. In addition, the stability of the result should be verified despite the changed presentation sequence (see Table 5) and whether the difference between blocks 1 and 3 is due to the positioning of the blocks or the sensitive item. The assumptions, hypotheses, and theoretical conceptualizations are the same as in Study 2, adding that I wanted to test any position effect, examine the stability of the results, and avoid the problem of the repeated measurement.

Study 3

Study 3 was conducted to verify and confirm the results of Study 2 and take into account the aforementioned limitations and problems.

Method

Participants and Procedure

I obtained the data from an online survey conducted in May 2013. The participants were recruited via the mailing list of the University of Gießen. The sample comprised 948 participants, 579 were female (47.3% had completed upper secondary education). The age ranged from 18 to 77 years (M = 31.12, SD = 11.17).

I adopted the design and procedure of Study 2. The survey was randomly split into two conditions (see Table 5). The experimental condition (N = 474) received the sensitive item which was presented prior to the nonsensitive items. The control condition (N =474) was presented with the sensitive item after the nonsensitive items. In this study, block 1 and block 3 were switched. In both conditions, the questionnaire started with the block in which only nonsensitive items were presented. Also in this study, the ‘control’ item “I am angry that one can easily make purchases in gasoline stations if necessary” changes its position in the

different conditions. Hence, the block with the sensitive item "Jews have too much influence in the world" was placed in block 3 (see Table 5).

-Table 5 about here-

In Study 3, only hypotheses 1 and 3 were tested to demonstrate how the results in block 1 with only nonsensitive items will be, in order to corroborate the results and the hypothesis of Study 2. Block 1 was, in Study 2, located at the end of the questionnaire and thus influenced by a sequence effect. Hypothesis 2 is not included in the analysis because the block in which the hypothesis was tested is, in this study, at the end of the questionnaire and thus also distorted by the repeated measurement.

H1: The agreement to the nonsensitive items changes if a sensitive item is included.

→ Hypothesis was tested in block 1.

H3: The indication of the number of *yes* answers leads to distortions.

→ Hypothesis was tested in block 1.

Results: Item difficulty

H1: The agreement to the nonsensitive items changes if a sensitive item is included.

To corroborate the finding in block 1 in Study 2 and to show that the biased item difficulty affected only the block with the sensitive item and not the block with only nonsensitive items, a t-test in block 1 (only nonsensitive items) was applied. Furthermore, the possible repeated measurement or learning effect problem should be excluded because the block with only nonsensitive items was located at the beginning of the questionnaire. Again, the mean levels were calculated using the sum of the direct *yes* answers to four nonsensitive

items. The 'control' item "I am angry that one can easily make purchases in gasoline stations if necessary" was not included. The test showed that the mean levels between experimental and control condition do not differ significantly, t(928) = .805, p = .42 (experimental condition M = 1.22, SD = .85; control condition M = 1.27, SD = .87).

Results: Biased individual aggregate value

H3: The indication of the number of yes answers leads to distortions.

In order to find out if the increased individual aggregate value only occurred in the block with the sensitive item and to test whether the results of the block with only nonsensitive items are a product of the repeated measurement and learning effect problem, a 2 x 2 design and repeated measures ANOVA was calculated for the first block with only nonsensitive items. The mean level of the sum of direct *yes* answers was generated from all five items in block 1. The main effect in block 1 was significant, $F(1, 635) = 76.13$, $p = .000$, partial $\eta^2 = 0.10$, confirming (see Figure 3) that the mean level of the individual aggregate value (M = 1.76, SD = 1.46) is significantly higher than the mean of the sum of direct *yes* answers (M = 1.39, SD = .87).

-Figure 3 about here-

Discussion

The H1 can be confirmed again with the results of Study 3. The study underlines the findings of Study 2 without the problem of repeated measurement bias. In block 1, which was presented at the beginning of the questionnaire and included only nonsensitive items, no effect was found and the results were constant. These results are in line with my

conceptualization that the sensitive item is at least partly responsible for the distortion. The approval rate of the nonsensitive items increased if a sensitive item was included.

The plausible results found for the third hypothesis in Study 2 could not be confirmed in Study 3 when the repeated measurement was controlled. However, after controlling the repeated measurement problem in Study 3, the result in the nonsensitive item block changed. Here, I found a significant difference between individual aggregate value and the mean of the sum of direct *yes* answers in block 1 (only nonsensitive items). As assumed, the individual aggregate value increased also in the nonsensitive item condition. This result is a further indication that the list experiment is not working in the way it was conceptualized. In sum, the results confirm the hypothesis that the individual aggregate value is distorted. In both blocks, the individual aggregate value is biased upwards and suggests that the list experiment generated no valid results.

6 General discussion

The main goal of the studies was to ascertain different moderators in order to show why the list experiment provides inconsistent results. I combined qualitative and quantitative methods to create an adequate basis for the analysis. In total, the results of three studies provide some answers regarding the possible difficulties with the list experiment.

First, the cognitive interviews have demonstrated that there were no problems of understanding in the list experiment. The introduction of the list experiment was mostly understood by the respondents and was judged as easy. These results are also supported by those presented by Droitcour et al. (2004) via cognitive interviews and by Coutts and Jann (2011) who implemented a direct self-report question in their questionnaire. The authors could show that 91.8 percent of the respondents understood the introduction of the list experiment completely. In addition, I was able to report that only a few respondents noticed

the sensitive item and were able to distinguish between the sensitive and nonsensitive item. However, since the nonsensitive items were described by the respondents as very easy to answer because of personal relevance, a shift in item difficulty or a shift of the response pattern due to the sensitive item seems plausible. Therefore, the quantitative experimental Studies 2 and 3 try to give more precise findings regarding the possible shifted item difficulty. Studies 2 and 3 were able to corroborate the hypothesis that the agreement to the nonsensitive items was changed if a sensitive item had been included. The approval rate of the nonsensitive items increased significantly when the sensitive item was placed prior to the nonsensitive items. By using an experimental design, the assumption from the cognitive interviews (Study 1) could be confirmed: a shift in item difficulty occurs. It is possible that in the experimental condition in the list experiment, a higher mean level is achieved because of the shifted item difficulty and the associated greater agreement to the nonsensitive items. Accordingly, the increase in the mean level of the test condition in the list experiment might not be attributed to the agreement to the sensitive item. A further factor which is responsible for possible distortions or provides an explanation for the inconsistent results is the indication of the number of *yes* answers (i.e., the individual aggregate value). The data support this hypothesis. It has been shown that the individual aggregate value increased also in the condition with only nonsensitive items when the problem of the repeated measurement had been taken into account. The individual aggregate value is significantly higher in both conditions (sensitive item included and nonsensitive items only) compared to the sum of *yes* answers in direct self-reports. This result suggests that in the list experiment, the indication of the number of *yes* answers (individual aggregate value) is biased upwards. This type of distortion occurred in the list experiment not only in the test condition but also in the baseline condition in which, due to the nonsensitive items, the individual aggregate value and the sum of *yes* answers of the direct self-report questions should be similar or equal. On the basis of

the upward biased individual aggregate value, another factor for the inconsistent results of the list experiment can be considered at this point. This result resembles those found in a study by Gosen et al. (in prep.) in the broadest sense. Here, the authors tested whether the mean level in the list experiment increased due to the number of items. For the analysis, four nonsensitive items were compared versus a condition consisting of five nonsensitive items. The authors found a significant difference between the two conditions. According to the logic of the list experiment, this result should not have occurred if the increased mean level in the test condition was attributed only to the sensitive item. The distortion caused by the different number of items in the baseline and test condition is a further determinant that calls the validity of the list experiment into question.

The results of the present studies are subject to some limitations. One main limitation is the biased repeated measurement. Due to this problem, it is not valid to obtain more information from the studies. For further analyses, it would be more useful to combine Studies 2 and 3 into one study. Thus, the lack of randomization is no longer a problem and one would receive comparable data. However, a disadvantage of this procedure is that the dataset has to be very large because at least six conditions (includes two sensitive items, e.g., anti-Semitism and Islamophobia) should be tested.

Furthermore, the results clearly show that the valence of the sensitive item plays a substantial role toward the nonsensitive items. Tsuchiya, Hirai, and Ono (2007) indicated, in a study of properties of the item count technique, that sensitive and nonsensitive items should have a similar valence. However, the idea on which this is based was that the respondents are not suspicious of the list experiment and not, as I assumed, that the sensitive item changed the valence of the nonsensitive items and their related approval rate. Nevertheless, the aspect of the equivalence of the items should be examined in further studies. It should be tested whether an equal valence yields an aforementioned effect or whether other possibilities of

distortions appear. Especially in the field of prejudice research, I assume that Tsuchiya et al.'s (2007) suggestion of the same valence of the item is difficult to implement. If the valence of the items would be similar, the risk of a floor or ceiling effect is very high.

The reported studies evidenced again that the validity of the list experiment is influenced or affected by several factors that occurred independently of one another or together. When these results are taken into account, the inconsistent findings of the literature are not surprising anymore. In these studies and in a previous study (Gosen et al. in prep.), it has emerged that the following moderating factors determined the validity of the list experiment:

1. The various numbers of items in the different conditions biased the results of the individual aggregate value.
2. The item difficulty of the nonsensitive items is biased if a sensitive item is included.
3. In general, the individual aggregate value is biased upwards.

According to the present findings, it is questionable whether the list experiment should be used in the field of prejudice or in other sensitive fields of research. If there is a need to control for social desirability biases, it is crucial to employ an instrument that does not introduce further distortions. The list experiment cannot provide this functionality. I conclude that the findings are mainly consequences of an inaccurate, thus again biased, cognitive process in individuals. I assume this distortion to occur on such a general level that it cannot be coped with empirically by introducing marginal conditions to the list experiment.

ACKNOWLEDGMENT

The author thank Uli Wagner, Peter Schmidt, Stefan Thörner and the members of the graduate school `Group Focused Enmity` for their helpful comments on an earlier draft of this paper. The author would also like to thank Lisa Trierweiler for the English proof of the manuscript. This work was supported by the Interdisciplinary Research Training Group 'Group Focused Enmity' located in Bielefeld and Marburg, Germany, funded by the Deutsche Forschungsgemeinschaft (GRK 884).

References

Ahart, A. M., Sackett, P. R.: A new method of examining relationships between individual difference measures and sensitive behavior criteria: evaluating the unmatched count technique. Organ. Res. Methods, **7**, 101–114 (2004)

Biemer, P., Jordan, B. K., Hubbard, M. L., Wright, D.: A test of the item count methodology for estimating cocaine use prevalence. In: Kennet, J., Gfroerer, J. (eds.) Evaluating and Improving Methods Used in the National Survey on Drug use and Health, Substance Abuse and Mental Health Service Administration, pp. 149–174. Office of Applied Studies, Rockville (2005)

Blair, G., Imai, K.: Statistical analysis of list experiments. Pol. Analys. , 2**0**, 47–77. (2012)

Chaudhuri, A., Christofides, T. C.: Item Count Technique in estimating the proportion of people with a sensitive feature. J. Stat. Planning Infer. **137**, 589–593 (2007)

Coutts, E., Jann, B.: Sensitive questions in online surveys: experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). Sociol. Methods Res. **40**, 169–193 (2011)

Dalton, D. R., Wimbush, J. C., Daily, C. M.: Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. Pers. Psychol. **47**, 817–828 (1994)

Dalton, D. R., Daily, C. M., Wimbush, J. C.: Collecting "sensitive" data in business ethics research: a case for the unmatched count technique (UCT). J. Bus. Ethics **16**, 1049–1057 (1997)

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsely, T. L., Visscher, W., Ezzati, T. M.: The item count technique as a method of indirect questioning: a review of its development and a case study application. In: Biemer, P., Groves, R. M., Lyberg, L., Mathiowetz, N., Sudman, S. (eds.) Measurement Errors in Surveys, pp. 185–210. Wiley, New York (1991)

Gilovich, T., Griffin, D., Kahneman, D.: Heuristics and biases: the psychology of intuitive judgment. Cambridge University Press, United Kingdom (2002)

Gosen, S., Schmidt, P., Thörner, S., Leibold, J.: Is the list experiment doing its job? Inconclusive evidence! (in prep.)

Greenwald, A. G., McGhee, D. E., Schwartz, J. L. K.: Measuring individual differences in implicit cognition: the implicit association test. J. Pers. Soc. Psychol. **74**, 1464–1480. (2010)

Heitmeyer, W.: Deutsche Zustände, Folge 1.[German states. Vol. 1]. Suhrkamp, Frankfurt am Main (2002)

Holbrook, A.L., Krosnick, J. A.: Social desirability bias in voter turnout reports: tests using the item count technique. Publ. Opin. Q. **74**, 37–67 (2010)

Jones, E. F., Forrest, J. D.: Underreporting of abortion in surveys of U.S. women: 1976 to 1988. Demogr. **29**, 113-126 (1992)

Kane, J. G., Craig, S. C., Wald, K. D.: Religion and presidential politics in Florida: a list experiment. Soc. Science Q. **85**, 281–293 (2004)

Karlan, D., Zinman, J.: List randomization for sensitive behavior: an application for measuring use of loan proceeds. J. Dev. Eco. **98**, 71–75 (2012)

Kuklinski, J. H., Cobb, M. D., Gilens, M.: Racial attitudes and the "New South". J. Polit. **59**, 323–349 (1997a)

Kuklinski, J. H., Sniderman, P. M., Knight, K., Piazza, T., Tetlock, P. E., Lawrence, G. R., Mellers, B.: Racial prejudice and attitudes toward affirmative action. Am. J. Polit. Science **41**, 402–419 (1997b)

Krumpal, I.: Estimating the prevalence of xenophobia and anti-Semitism in Germany: a comparison of randomized response and direct questioning. Soc. Science Res. **41**, 1387–1403 (2012)

Krumpal, I.: Determinants of social desirability bias in sensitive surveys: a literature review. Qual. Quant. **47**, 2025–2047 (2013)

Krysan, M., Couper, M. P.: Race in the live and the virtual interview: racial deference, social desirability, and activation effects in attitude surveys. Soc. Psychol. Q. **66**, 364–383 (2003)

Miller, J. D.: A new survey technique for studying deviant behavior. Unpublished doctoral dissertation, George Washington University (1984)

Paulhus, D. L.: Two-component models of socially desirable responding. J. Pers. Soc, Psychol. **46**, 598–609 (1984)

Rayburn, N. R., Earleywine, M., Davison, G. C.: Base rates of hate crime victimization among college students. J. Interpers. Violence **18**, 1209–1221 (2003)

Roese, N. J., Jamieson, D. W.: 20 years of bogus pipeline research—a critical-reviewandmetaanalysis. Psychol. Bull. **114**, 363–375 (1993)

Sirken, M.: Household surveys with multiplicity. J. Am. Stat. Assoc. **65**, 257–266 (1970)

Sniderman, P. M., Grob, D. B.: Innovations in experimental design in attitude surveys. Ann. Rev, Soc. **22**, 377–399 (1996)

Streb, M. J., Burrell, B., Frederick, B., Genovese, M. A.: Social desirability effects and support for a female American president. Publ. Opin. Q. 72, 76–89 (2008)

Tourangeau, R., Rips, L. J., Rasinski, K. A.: The Psychology of Survey Response. Cambridge University Press, Cambridge (2000)

Tourangeau, R., Yan, T.: Sensitive questions in surveys. Psychol. Bull. **133**, 859–883 (2007)

Tsuchiya, T., Hirai, Y., Ono, S.: A study of the properties of the item count technique. Publ. Opin. Q. **71**, 253–272 (2007)

Tversky, A., Kahneman, D.: Judgment and uncertainly: heuristics and biases. *Publ. Science* **185**, 1125–113 1(1974)

Warner, S. L.: Randomized response: a survey technique for eliminating evasive answer bias. J. Am. Stat. Assoc. **60**, 63–69 (1965)

Willis, G. B.: Cognitive interviewing: a tool for improving questionnaire design. Sage, Thousand Oaks (2005)

Willis, G. B., DeMaio, T. J., Harris-Kojetin, B.: Is the bandwagon headed to the methodological promised land? Evaluating of the validity of cognitive interviewing techniques. In: Sirken, M. G., Hermann, D. Schechter, J. S., Schwarz, N., Tanur, J., Tourangeau, R. (eds.) Cognition and survey research , pp. 133–154. Wiley, New York (1999)

Wimbush, J. C., Dalton, D. R.: Base rate for employee theft: convergence of multiple methods. J. Appl. Psychol. **82**, 756–763 (1997)

Zick, A., Wolf, C., Küpper, B., Davidov, E., Schmidt, P., Heitmeyer, W.: The syndrome of group-focused enmity: the interrelation of prejudices tested with multiple cross-sectional and panel data. J. Soc. Issues **64**, 363–383 (2008)

Zimmerman, R. S., Langer, L. M. Improving estimates of prevalence rates of sensitive behaviors: the randomized lists technique and consideration of self- reported honesty. J. Sex Res. **32**, 107–117 (1995)

Figure 1. Mean level of the sum of direct *yes* answers and of the individual aggregate value in block 1 (Study 2).

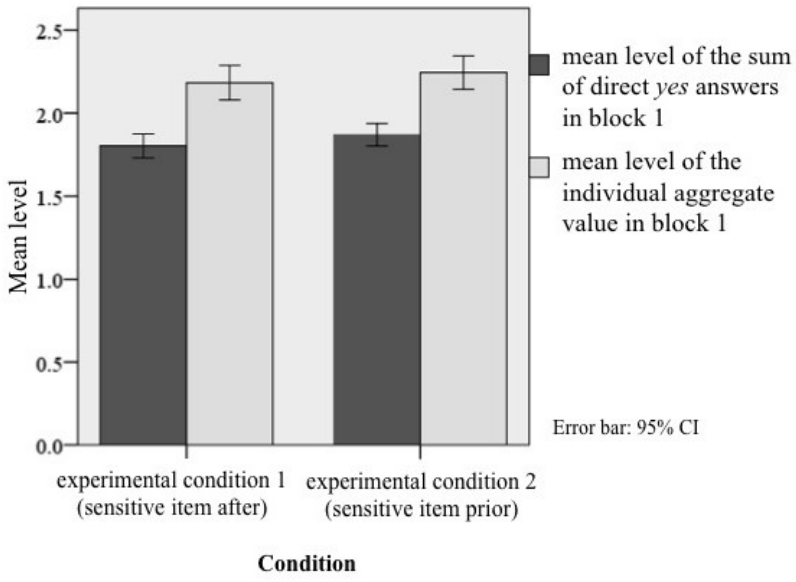


Figure 2. Mean level of the sum of direct *yes* answers and of the individual aggregate value in blocks 1 and 3.

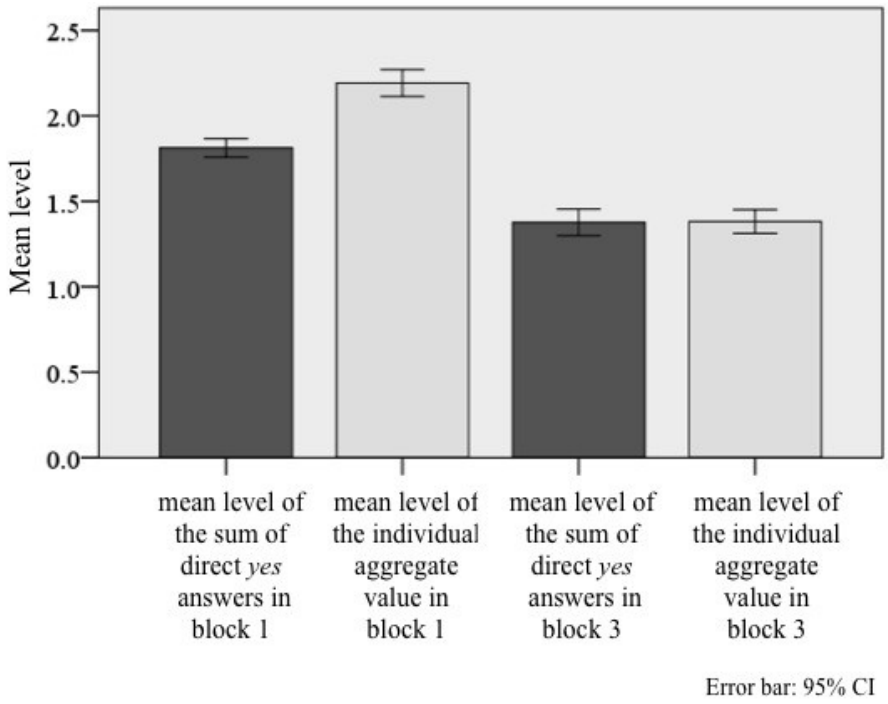


Figure 3. Mean level of the sum of direct *yes* answers and of the individual aggregate value in block 1.

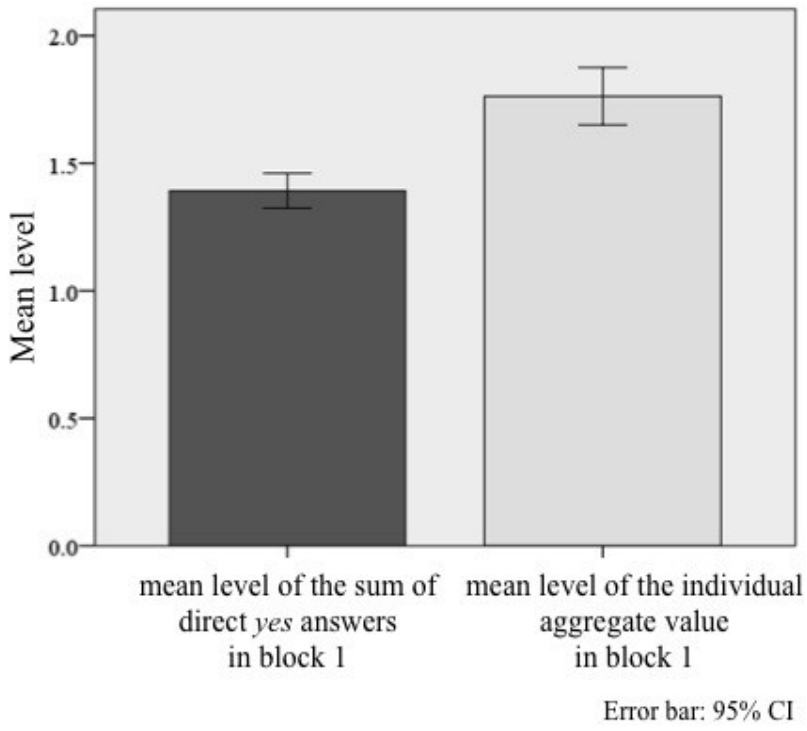


Table 1 Overview of the different studies.

Study	**Design**	**N**
Study 1	Cognitive interviews	N = 7
Study 2	Experimental design with three different segments	N = 1,878
Study 3	Experimental design with three different segments (two blocks were changed)	N = 948

Table 2 Study 1, sequence within the questionnaire.

Sequence	**Questionnaire 1**	**Questionnaire 2**	**Questionnaire 3**	**Questionnaire 4**
Sequence 1	GFE-Syndrome items	List experiment with sensitive item about "Jews"	GFE-Syndrome items	List experiment with sensitive item about "Muslims"
Sequence 2	List experiment with sensitive item about "Jews"	GFE-Syndrome items	List experiment with sensitive item about "Muslims"	GFE-Syndrome items

Notes: Questionnaire 1 included two respondents, Questionnaire 2 included one respondent, Questionnaire 3 included two respondents, and Questionnaire 4 included two respondents.

Table 3 Study 2, experimental and control condition (sequence within the questionnaire)

Condition	**Block 1**	**Block 2 (distractor)**	**Block 3**
Experimental (IV)	One *sensitive* item; Four nonsensitive items (DV); Indication of the number of *yes* answers (DV)	One sensitive item; Four nonsensitive items	One *nonsensitive* item; Four nonsensitive items (DV); Indication of the number of *yes* answers (DV)
Control (IV)	Four nonsensitive items; One *sensitive* item (DV); Indication of the number of *yes* answers (DV)	Four nonsensitive items; One sensitive item	Four nonsensitive items; One *nonsensitive* item (DV); Indication of the number of *yes* answers (DV)

Notes: Each item was presented on a separate page. In the experimental condition the sensitive and one of the two possible nonsensitive items was presented prior to the four nonsensitive items. IV = independent variable, DV = dependent variable.

Table 4 Nonsensitive items in the different blocks.

Block 1	**Block 2**	**Block 3**
1. I am angry that professional athletes are getting million dollar plus salaries.	1. I am angry that so many large companies pollute the environment.	1. I am angry that ARD and ZDF spend millions of Euros for the rights of football broadcasting.[+]
2. I am angry that the costs of gasoline regularly increase at Easter time and in the school holidays.	2. I am angry that the federal government increased Hartz 4.*	2. I am angry that more is done for renewable energies.
3. I am angry that seat belts are required to be used when driving.	3. I am angry that using a mobile phone while driving is not allowed.	3. I am angry that there is no speed limit on German highways.
4. I am angry that the German railroad has so many delays.	4. I am angry that the practice fee is cancelled.	4. I am angry about the way gas prices keep going up.
5. I am angry that Jews having too much influence in the world.	5. I am angry that Muslims having too much influence in the world.	5. I am angry that one can easily make purchases in gasoline stations if necessary.

* Hartz 4 is the German unemployment assistance. [+]ARD and ZDF are the fee-financed public broadcasting services. In Block 3 the fifth item was the 'control item', which also changed the position as the sensitive item in block 1 and 2 in the experimental condition.

Table 5 Study 3, experimental and control condition (sequence in the questionnaire); blocks 1 and 3 change their positions.

Condition	**Block 1**	**Block 2 (distractor)**	**Block 3**
Experimental (IV)	One *nonsensitive* item; Four nonsensitive items (DV); Indication of the number of *yes* answers (DV)	One sensitive item; Four nonsensitive items	One *sensitive* item; Four nonsensitive items (DV); Indication of the number of *yes* answers (DV)
Control (IV)	Four nonsensitive items; One *nonsensitive* item (DV); Indication of the number of *yes* answers (DV)	Four nonsensitive items; One sensitive item	Four nonsensitive items; One *sensitive* item (DV); Indication of the number of *yes* answers (DV)

Notes: The random assignment in terms of gender, age, and education was successful across the conditions. Block 2 was the distractor block. IV = independent variable, DV = Dependent variable.

8. Final discussion

As has been mentioned previously in this dissertation, it is difficult to find a critical view of the list experiment within the published literature. The general goal of the list experiment is to reduce the socially desirable response bias by guaranteeing respondents privacy and anonymity. Thus, the list experiment is able to create an estimate of the proportion of people who were angry about the sensitive item. In order to determine the social desirable bias the estimation of the list experiment is compared to direct self-report questions. If there is a social desirable bias, the estimate of the list experiment should be higher than the direct self-report. However, in the literature there are some studies in which the list experiment did not produce higher estimates than direct self-report questions (Droitcour et al, 1991; Ahart & Sackett, 2004; Biemer et al., 2005) and it is not apparent which factors lead to this failure. The question why the list experiment failed even with high expenditure and detailed planning and, in contrast, provided expected results in other studies without a high degree of preparation, needs to be answered in order to assess its validity. The reasons for its ineffectiveness are based merely on assumptions, which researchers try to explain. Hence, there is a lack of sufficient analyses that detect the inconsistent results, the effectiveness of the list experiment, and the appropriate factors that might be responsible for the failure of this technique. The aim of this dissertation was to make proper propositions about its validity, consistency, and to find specific marginal factors/moderators that determine its ineffectiveness.

The present research was able to prove the inconsistent results in the field of prejudice research (Manuscript #1) and could find specific factors/moderators that are responsible for the failure or rather the inconsistent results of the list experiment and explain where the difficulties of this technique are (Manuscript #1 and Manuscript #2).

In Manuscript #1 the list experiment could not provide valid results. While in Study 1 (representative) the list experiment received results in the expected direction and produced a higher estimate of the proportion of people who were angry about the sensitive item "Jews have too much influence in the world" than the direct self-report question, in Study 2 (representative) – which was a modified repetition of Study 1 – the list experiment failed completely. In this study, the list experiment showed no significant difference between baseline and test condition, and the direct self-report item yielded a higher estimate than the list experiment. Accordingly, the list experiment received a smaller estimate than the direct self-report question and thus it was not able to detect socially desirable response bias. The different results of Studies 1 and 2 questioned its validity strongly. In order to test it in more detail and to find a possible factor of its failure, a further study was conducted.

The third study in this manuscript (Study 3) was conducted as an online survey, which was not representative, and it implemented three list experiments. Two of them contained two different prejudice items (anti-Semitism and Islamophobia), and the other one contained only nonsensitive items. However, no consistent results could be found. The first list experiment in the online study (Study 3) that studied anti-Semitism yielded almost the same invalid results as Study 2. Here, the list experiment also did not receive a significant mean level difference. Like in Study 2, the estimate of the direct self-report question also was not significantly higher than the list experiment. It did, however, show a tendency in the assumed direction. Again, the list experiment has not revealed the assumed social desirability induced underestimation of prejudices. In contrast, the list experiment with the Islamophobic sensitive item obtained a significant mean difference in the expected direction; it also indicated a higher estimate than the direct self-report question, however, this difference was not significant. From this perspective, the list experiment proceeded in the assumed direction, but

again, it was not able to yield clear and significant results. It can therefore be said that the list experiment was able to indicate a tendency of socially desirable response bias.

The third list experiment in this study (Study 3) checked its validity and also had the purpose of finding a factor that might explain the inconsistent results or provide a reason for the failure of the list experiment. This factor in question was the number of items, i.e. whether the change or the increase in the mean occurred because of the higher number of items in the test condition. As mentioned before, in this condition the experiment consisted of nonsensitive items only. Hence, four nonsensitive items were compared to five nonsensitive items. The neutrality of the fifth nonsensitive item was tested via a frequency analysis because it was requested as direct self-report question in a pretest survey. Here, only 2 of 75 respondents showed anger about this item. The analysis revealed a significant mean difference between baseline and test condition. This result implies enormous consequences for the validity of the list experiment itself because it indicated that the higher number of items could be responsible for the higher mean level in the test condition. In other words, the increase of the mean in the test condition depends not only on the content of the particular items but also on the number of items.

The final study, Study 3, consisted of two waves to test intraindividual stability. Here, the respondents showed that their responses were quite stable when the baseline condition included only four items. When there were only four nonsensitive items in the baseline condition compared to five nonsensitive items, the probability of the respondents giving the same answer twice was significantly higher in the baseline condition. This result also supports the assumption that the number of items in the different conditions is a factor that causes possible distortion in answering the list experiment. In sum, Manuscript #1 was able to show that the evaluation of the efficiency of the list experiment results in inconclusive

evidence. Consequently, the list experiment could not detect or avoid social desirability responses.

Manuscript #2 is linked to the research focus of the validity of the list experiment and was able to find various factors (moderators) that can partly explain the inconsistent results. Study 1 (see Chapter 4, Table 5 for an overview of the studies) used qualitative interviews in which the list experiment was read to the respondents. Most of them did not have any problems understanding it. Misunderstanding, therefore, can be excluded as a possible source of inconsistent results[1]. Furthermore, in only three of seven interviews did the respondents recognize the sensitive item. They did not, however, classify the sensitive item as negative, difficult or outrageous question. The other four respondents also did not perceive this sensitive question as outlier. Nevertheless, the nonsensitive items in the list experiment were noticed as items that were easier to answer because of their included personal and everyday life issues. As a consequence of these findings, the hypothesis that the sensitive item changed the agreement to the nonsensitive items (shifted item difficulty) was tested in quantitative experimental studies.

Study 2 (online experimental study) could confirm this hypothesis and indicated that the respondents agreed significantly more to the nonsensitive items when the sensitive item was included. For the list experiment, this result means that the mean level in the test condition increased due to the shift in item difficulty and not due to the content of the sensitive item, as the list experiment presupposes. A further factor (moderator) that might lead to a possible distortion within the list experiment is the counting that respondents had to perform. They were asked to indicate the number of items that they agreed to (this value is referred as individual aggregate value).[2] However, the hypothesis that the position of the

[1] Droitcour et al. (1991) and Coutts & Jann (2011) assumed and tested it too.
[2] In the questionnaire the question to indicate the number of items was: "Please indicate how many of the previous questions you answered with *yes*."

sensitive item had an effect on the indication of the number of *yes* answers was not corroborated in Study 2. Furthermore, the hypothesis that the individual aggregate value lead to distortions or deferred in general could not be supported by this study. Here, the individual aggregate value within the sensitive and nonsensitive item condition was as expected. The individual aggregate value was significantly higher in the condition with the sensitive item compared to the sum of the direct *yes* answers.[3] In the condition with the nonsensitive item, the individual aggregate value and the sum of direct *yes* answers did not differ significantly. On the first glance, this finding seemed to be able to confirm the effectiveness of the list experiment or the functioning of the individual aggregate value. In other words, by following the logic of the list experiment, it is accurate that the individual aggregate value provided higher estimates in the sensitive item condition than in the nonsensitive item condition.

Study 3 (experimental online study) tested a possible position/sequence effect due to the different positions of the segments in Study 2. In addition, the study tested also the stability of the results of Study 2 and tried to avoid the problem of the repeated measurement. In this study hypotheses were the same as in Study 2. In Study 3, two segments were turned in their order – the nonsensitive item condition was placed at the beginning of the questionnaire, which excludes the possibility of distortion by a sequence effect. The first hypothesis that the agreement to the nonsensitive items changed when a sensitive item is included, could be substantiated in Study 2 within the nonsensitive item condition. Here, the two conditions did not differ significantly. Therefore, the increased mean in the test condition or rather the increased agreement to the nonsensitive item appeared only when a sensitive item was included. Study 3 could corroborate the third hypothesis that the procedure to indicate the number of *yes* answers is distorted. The individual aggregate value yielded significantly higher estimates than the sum of direct *yes* answers. This finding implies that

[3] Sum of direct *yes* answers is formed out of the sum of *yes* answers of the five direct-self report questions.

within the list experiment the indication of the number of items is biased in the baseline and test condition. Also in this case, the sensitive item is not exclusively responsible for the increased mean in the test condition.

In combining the findings of Manuscript #1 and Manuscript #2, I conclude that the list experiment did not provide valid and consistent results. In the above-mentioned studies, it could be shown that there are too many factors/moderators that lead to distortions, inconsistencies and that influenced the functionality of the list experiment. In sum, three moderating factors were found, as mentioned in Manuscript #2, "that occurred independently of one another or together." In the following, these moderators are briefly listed:

1. The various numbers of items in the different conditions bias the results of the individual aggregate value;
2. the item difficulty of the nonsensitive items gets biased if a sensitive item is included;
3. in general, the individual aggregate value is biased upwards.

It is not obvious which factor effectively affects the answering of the list experiment individually, but I presented sufficient evidence that the list experiment did not produce efficient results. Like the suspected publication bias assumes, the possible number of unsuccessful trials might be higher as the published studies indicate.

The findings of the presented dissertation suggest that the current execution of the list experiment is problematic and also implies further unidentified factors. These factors might be increased measurement errors, socially desirable responses or inconsistent results. In sum, I conclude that it is questionable whether the list experiment should be used in prejudice research or other sensitive research topics. As already mentioned in Manuscript #2, "[i]f there is a need to control for social desirability biases, it is crucial to employ an instrument that does not introduce further distortions. The list experiment cannot provide this functionality."

9. Outlook

As I demonstrated in the previous chapter, research has shown that the question of validity of the list experiment and the possible reasons of its inconsistent results could partly be answered. In the process, new questions appeared that imply that further research is required to examine the factors for the inconclusive evidence of the effectiveness of the list experiment.

In a next and, by now, unavoidable step, it is crucial to determine the assumed publication bias and to get an overview of the list experiment's inconsistent results through an extended meta-analysis. This analysis should also take into account unpublished manuscripts as well as published ones. Further, it might be useful to analyze studies that work with representative samples. A reason for this is that the meta-analysis from Tourangeau and Yan (2007: 873) showed that the list experiment in undergraduate samples have a tendency to produce higher estimates and in general population surveys or rather representative samples the list experiment is not able to generate higher estimates than the direct self-report questions. For this purpose, it is important to test whether the failure of the list experiment is linked to the representativity and the specific sensitive issue. The researchers Auspurg, Jann and Krumpal (2012) are currently performing and collecting data for a meta-analysis in the area of indirect methods. Hence, they are trying to test, among other aspects, the publication bias of the RRT (randomized response technique) and the list experiment.

Apart from that, the list experiment's functionality should also be tested within different groups. Tsuchiya, Hirai and Ono (2007), as well as Holbrook and Krosnick (2010) divided their respondents in different groups, for example of age, education, gender and race and tested how the estimates of the list experiment differed from direct-self report questions. However, in order to test the validity of the list experiment, as in Manuscript #1, and to test

the effect of sensitive and nonsensitive items, a comparison of extreme groups would be interesting. For this purpose, groups had to be used whose ideology is clearly obvious and whose willingness to express their ideology is much stronger than their need to respond in a socially desirable way. The basic procedure would be that the list experiment is conducted with appropriate sensitive items, e.g., anti-Semitism plus direct self-report questions. In this context, it should be tested whether a difference between indirect and direct question appears. The assumption is that, for instance, in right-wing extremist groups, the differences within the list experiment should correspond almost exactly to the agreement in the direct self-report question. The reason is that groups with strong ideologies have lower social desirability bias and do not want to adapt to the general social norm but to express their ideology. And if there should appear a significant difference between direct-self report question and the list experiment for these groups, e.g., even the direct-self report question would receive higher values than the list experiment, it might be assumed that there could be further factors that are responsible for the ineffectiveness of the list experiment. This would additionally undermine the list experiment's validity.

Furthermore, it might be useful to conduct an additional experimental design to investigate the mentioned factors/moderators in Manuscript #2. In this case the factor might be the mean of the test condition, which is affected by the number of items, as well as the cognitive processes the respondents use while answering the list experiment. As evidenced by Manuscripts #1 and #2, the mean of the test condition increased both because of the higher number of items and the shifted item difficulty of the nonsensitive items because of adding a sensitive item. Therefore, it is not mandatorily caused by the content of the items. In order to check in which way the respondent's answers are dependent on the number of items and on the predetermined response categories (from 0–4 items or 0–5 items), the concept of anchoring effects, which is embedded in the information processing theory, could be used.

Basically, it is a judgment heuristic in which the individual is oriented towards an optional 'anchor' during a decision making process (Tversky & Kahneman, 1974). The anchor within the list experiment could be the number in the introduction of the list experiment ("Could you please tell me how many of the following *four/five* statements you have also been angry about?") or the numbered response categories, which are presented directly after the items of the list experiment. In order to obtain an anchoring of a judgment, an anchor in form of a number should be set within each introduction of the list experiment it. One version could be that in the introduction and within the response categories of the list experiment the number five is presented, although there are actually only four items listed. By varying the number of the anchor, it could be tested whether the judgment of the participants might be influenced by the anchor "five" or "four" rather than by the content of the items. Subsequently, a scale of intuitive and rational processing (Rational - Experiential Inventory [REI]; Epstein et al., 1996) could be applied to consider the cognitive process of the respondents, which was discussed in Manuscript #2. The whole scale consists of two subscales. On the one hand, the Need for Cognition Scale (NC), which should measure the analytical-rational or also the systematic processing. On the other hand, the Faith in Intuition Scale (FI), which should provide data to heuristic or rather to experiential-intuitive processing (German version from Keller, Bohner, & Erb, 2000). Thus it could also be tested, for example, if persons with high values on the NC scale show fewer distortions by the number presented in the introduction (anchor). In other words, the wrong numbers in the introduction have less influence to the answers of the respondents.

In total and because of the critical validation of the list experiment it has been shown that indirect (unobtrusive) response methods do not lead mandatorily to more truthful and valid self-reports regarding sensitive issues. The point is that they partly provide other, not identifiable and misleading, problems. This difficulty includes not only the method that was

investigated in this dissertation (its application is relatively new) but rather refers to other, more precise, and better researched methods, like the RRT (randomized response technique). In this sense, Lensvelt-Mulders et al. (2005: 323) described it adequately when they mentioned that, "A thorough look at the literature on RRTs reveals that 35 years of research have not led to a consensus or a description of best practices."

Despite of the failed studies, inconsistent and inconclusive results of the list experiment and the critical view of the RRT, it is not longer justifiable to use explicit direct self-report questions to avoid social desirable response biases when investigating sensitive issues. Consequently, survey research is still faced with the challenge to develop alternative measures that are able to avoid social desirable response bias in an easy and reliable way. A further possibility in this research area is using techniques that are less susceptible of social desirable response bias by employing another level of questioning. One of the advantages of these techniques over the list experiment is that these techniques are able to generate individual scores. One provided opportunity coming from social psychology is the concept of implicit measures (e.g., Fazio & Olson, 2003). The main difference between direct self-report questions (explicit measures) and implicit measures is that these techniques do not have an introspective access to the measured constructs and thus reduce the conscious control of the response process (De Houwer, 2006). Essentially, it means that the attitude of respondents should be measured without questioning them directly and therefore to receiving less distorted responses. Respondents are unaware (without the person knows that their attitudes are being accessed) which attitudes, stereotypes etc. are being measured (Fazio & Olson, 2003: 303). Thus, the process of answering is on an unconscious and automatic level. Therefore, many researchers believe that the output to socially sensitive issues is not biased by social desirability as direct self-report questions are (e.g., Fazio & Olson, 2003; Butz &

Plant, 2009; Gawronski, 2009; Greenwald et al., 2009; see Huddy & Feldman, 2009 for a critical view on this issue).

A technique that is well-known in the field of implicit measures is the Implicit Association Test[1] (IAT), introduced by Greenwald, McGhee, & Schwartz (1998). Basically, the test measures the speed of responses to different objects when they are classified into negative and positive categories. In other words, "[t]he Participants' task is to categorize stimuli as they appear on the screen" (Fazio & Olson, 2003: 299). The procedure, in a nutshell, is as follows. There are two types of stimuli – words that are either positive or negative (e.g., joy, hurt), and pictures that can be classified into two categories (e.g., African American or White American). The categories are then paired in such a way that, for example, positive and African American stand together, or positive and White American; the second category would then be negative and White American, or negative and African American, respectively. Respondents are then asked to respond to both words and pictures and assign them to one of the paired categories. The time of answering will be faster when both categories are highly associated. For instance, respondents with negative associations toward Blacks have shorter reaction times when the categories are 'Black + negative'. The fast pairing of the two categories can be seen as implicit prejudice (Frantz et al., 2004; Greenwald et al., 2009).

In order to understand social desirability bias, researchers correlate the assumed unbiased implicit prejudice measures with assumed biased explicit prejudice measures. The lower the correlation, the higher the assumed social desirability bias (Gawronski, 2009: 144). In the literature some studies corroborated this prediction (Banse, Seise, & Zerbes, 2001; Nosek, 2005; Riketta, 2006). Another study that encouraged the use of implicit measures in

[1] Demo Tests and Research Information of the Implicit Association Test: https://implicit.harvard.edu/implicit/ (belongs to 'Project Implicit' and was founded as a multi-university research collaboration by Greenwald, Banaji and Nosek).

the field of prejudice research, which compared explicit and implicit attitudes, is presented by Baron and Banaji (2006). The goal of the study was to show the development of the implicit–explicit correlation of attitudes towards social groups (blacks) across several ages. For this purpose, the researchers compared the implicit and explicit attitudes in three different groups of age (6 years, 10 years, and adults). In sum, they found that implicit race attitudes emerged at an early age and were constant across the development. On the contrary, the direct self-report answers became more egalitarian the older the participants were. This result implies that in direct self-reports the social pressure and societal demand to avoid prejudiced attitudes increases with age. In other words, it suggests that social desirability responses increase in self-reports because the respondents learn to adapt to the social norm when they get older. Thus, there are aspects that moderate the implicit-explicit correlation (Hofmann, Gawronski, & Gschwendner, 2005; Hofmann, Gschwendner, & Schmitt, 2005; Nosek, 2005; 2007). One of these aspects is the motivation to respond without prejudice (Gawronski, LeBel, & Peters 2007).

A study that tried to predict social desirability responses with implicit measures is from Schlauch et al. (2009). On the basis of external and internal motivation to respond without prejudice, the authors could show that respondents with a high internal motivation also had a high implicit–explicit correlation even when intoxicated with alcohol. Evidently, they had a higher level of control over their responses. In contrast, participants with an external motivation showed less control over their responses and a higher level of prejudice attitudes in the implicit measure, even if they were in the placebo group, i.e. they falsely believed that they had got alcohol.

Furthermore, another study corroborates the utility of the implicit measures to reduce the social desirability bias. Nier (2005) tested the dissociation between explicit and implicit racial attitudes with the aid of the bogus pipeline technique. The bogus pipeline technique is a

kind of lie detector method, which tries to reduce social desirability responses (see Chapter 3.1). The study was able to show that a significant correlation between explicit (i.e. direct self-report questions) and implicit (IAT) measures of racial attitudes occurred when respondents believed that the researcher could identify if they had answered truthfully. In contrast, when the bogus pipeline technique was not used, no correlation was found between explicit and implicit measures. According to the author, "the results suggest that as the motivation to report explicit attitudes that are consistent with implicit attitudes increases, the implicit–explicit relationship strengthens due to changes in self-reported explicit attitudes (implicit attitudes were not influenced by the information that participants received about the IAT)." (Nier, 2005: 49)

In addition, a meta-analysis from Greenwald et al. (2009) demonstrated that in measuring racial behaviors the IAT showed significantly higher predictive validity than direct self-report questions.

Despite the presented studies that facilitate the use of implicit measures to reduce the social desirability response bias, this research field and especially the relation between explicit and implicit attitudes is still under discussion (Gawronski & Conrey, 2004; Gawronski & Bodenhausen, 2006; Nosek, Greenwald, & Banaji, 2007; Gawronski, 2009).

There are many studies that failed to find effects of social desirability by comparing explicit and implicit measures (Dovidio, Kawakami, & Gaertner, 2002; Hofmann, Gawronski, & Gschwendner, 2005; Fishbein & Ajzen, 2010). As mentioned before, researchers assume that implicit measures are less susceptible for response factors such as social desirability. Egloff and Schmukle (2003) tried to find an explanation for the low correlations between implicit IAT measures and explicit self-report measures, especially in the area of anxiety (e.g., an overview on web-based IATs given by Egloff and Schmukle (2003) showed a correlation between implicit and explicit of $r = 0.24$ (Nosek, Banaji, &

Greenwald, 2002), correlation of self-esteem $r = 0.21$ (Bosson, Swann, & Pennebaker, 2000), correlation of anxiety $r = 0.14$ (Egloff & Schmukle, 2002)). They assumed that in the case of anxiety the socially desirable bias might be a moderator of the relation between implicit (IAT) and explicit (direct self-report questions) measures. The authors expected that anxiety would be slightly socially undesirable and therefore respondents with high social desirability values would show a lower correlation between implicit and explicit measures. The results from the two studies indicated that social desirability did not moderate the correlation between implicit and explicit measures in the area of anxiety.

A further study from Cunningham et al. (2001) compared IAT measures with scores of the Modern Racism Scale (McConahay, 1986). Here, the correlation between implicit and explicit measures were very low and ranged between $r = .08$ and $r = .26$.

Furthermore, Hofmann, Gawronski and Gschwendner (2005) tested the relationship between explicit self-report questions and the IAT in a meta-analysis. They reported for the correlation between implicit and explicit measures across 126 studies a mean effect size of .24. Within the meta-analysis the researchers investigated if social desirability (moderator) predicts the correlation between implicit and explicit measurement (study correlation) negatively. The authors found that there was a nonsignificant negative relationship between social desirability and 151 study correlations. In addition, they could show when three predictors were used simultaneously (spontaneity, introspection and social desirability) that social desirability predict the correlation between the IAT and the direct self-report question significantly in a positive direction and not as theoretically assumed in a negative direction.

Nevertheless, I would consider implicit measures might be a promising method to receive more valid and truthful responses from respondents especially in the area of prejudice research.

This dissertation has shown that the state-of-the-art methods to cope with social desirability response biases in general and the list experiment in particular are far from perfect, and that there are still pitfalls and non-ambiguous findings. Therefore, it indicates that more work needs to be done to tackle the problem of social desirability in survey research. The positive implications of improved methods are obvious – the more we know about the response bias that social desirability causes, the better get our chances to cope with it and improve research and data quality. Thus, this work made its scientific contribution by increasing our knowledge about the complex puzzle of social desirability bias as well as the methods that should reduce it.

10. References

Ahart, A. M., & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods, 7*, 101–114.

Auspurg, K., Jann, B., Krumpal, I., & von Hermanni, H. (2012). Randomized-Response-Technik: Hope or Hype? Eine Meta-Analyse unter Berücksichtigung von Publication-Bias [Randomized-response-technique: Hope or hype? A meta-analysis in consideration of publication bias]. *Paper presented at the First Mini-Conference of the Center of Quantitative Methods of the University of Leipzig. Asking Sensitive Questions: Theory and Data Collection Methods.*

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48,* 145–160.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science, 17*, 52–58.

Biemer, P. P., Jordan, B. K., Hubbard, M., & Wright, D. (2005). A test of the item count methodology for estimating cocaine use prevalence. In Kennet, J., and J. Goefrer (Eds.), *Evaluating and improving methods used in the national survey on drug use and health (DHHS Publication No. SMA 05-4044, Methodology Series M-5)* (pp. 149-174). Rockville, MD: Dept. of Health and Human Services Administration, Office of Applied Studies.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: the blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631–643.

Butz, D. A., & Plant, A. (2009). Prejudice control and interracial relations: The role of motivation to respond without prejudice. *Journal of Personality, 77*, 1311–1342.

Coutts, E., & Jann, B. (2011) Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods and Research, 40,* 169–193.

Cunningham, W. A., Preacher, K. J. & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 121*, 163–170.

De Houwer, J. (2006). What are implicit measures and why are we using them. In R. W. Wiers, and A. W. Stacy (Eds.), *The Handbook of Implicit Cognition and Addiction* (pp. 11-28). Thousand Oaks, CA: Sage Publishers.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82,* 62–68.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 185–210). New York: Wiley.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of Personality and Social Psychology, 83*, 1441–1455.

Egloff , B., & Schmukle, S. T. (2003). Does social desirability moderate the relationship between implicit and explicit anxiety measures? *Personality and Individual Differences, 35*, 1697–1706.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology, 7*, 390–405.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, 54, 297–327.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach. New York*. Taylor & Francis.

Frantz, C., Cuddy, A. J., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the computer: The race implicit association test as a stereotype threat experience. *Journal of Personality and Social Psychology Bulletin, 30,* 1611–1624.

Gawronski, B. (2002). What does the implicit association test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental Psychology, 49,* 171–180.

Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology, 50*, 141–150.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692–731.

Gawronski, B, & Conrey, F. R. (2004). Der Implizite Assoziationstest als Maß automatisch aktivierter Assoziationen: Reichweite und Grenzen [The implicit association test as a measure of actived associations: Scope and limits]. *Psychologische Rundschau, 55,* 118–126.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. Perspectives on Psychological Science, 2, 181–193.

Greenwald A.G., McGhee D.E., & Schwartz J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E., & Banaji, M.R. (2009). Understanding and using the implicit association test: III. meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31,* 1369–1385.

Hofmann, W., Gschwendner, T., & Schmitt, M. (2005). On implicit-explicit consistency: The moderating role of individual differences in awareness and adjustment. *European Journal of Personality, 19*, 25–49.

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly, 74*, 37–67.

Huddy, L., & Feldman, S. (2009). On assessing the political effects of racial prejudice**.** *Annual Review of Political Science, 12,* 423–47.

Keller, J., Bohner, G., & Erb, H.-P. (2000). Intuitive und heuristische Verarbeitung - verschiedene Prozesse? Präsentation einer deutschen Fassung des „Rational-Experiential Inventory“ sowie neuer Selbstberichtskalen zur Heuristiknutzung [Intuitive and heurisitc processing- different processes? Presentation of a German version of the “Rational-experiential inventory” as well as new self-reported scales for heuristic use]. *Zeitschrift für Sozialpsychologie*, 31, 87–101.

Lensvelt-Mulders, G. J. L. M., Hox, J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research. *Sociological Methods and Research, 33,* 319–348.

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination and racism* (pp. 91– 126). New York: Academic.

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations, 8,* 39–52.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134,*565–584.

Nosek, B. A. (2007). Implicit-explicit relations. *Association for Psychological Science, 16*, 65–69.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics, 6*, 101–115.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). Psychology Press.

Riketta, M. (2006). Gender and socially desirable responding as moderators of the correlation between implicit and explicit self-esteem. *Current Research in Social Psychology, 11,* 14–28.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859–883.

Tsuchiya, T., Hirai, Y., & Ono, S. (2007). A study of the properties of the item count technique. *Public Opinion Quarterly, 71*, 253–272.

Tversky, A., & Kahneman, D. (1974). Judgment and uncertainly: Heuristics and biases. *Public Science, 185*, 1125–1131.

Schlauch, R. C., Lang, A. R., Plant, E. A., Christensen, R., & Donohue, K. F. (2009). Effect of alcohol on race-biased responding: The moderating role of internal and external motivations to respond without prejudice. *Journal of Studies on Alcohol and Drugs, 70*, 328–336.

Appendix A: Content of enclosed CD-Rom

Questionnaire and transcription of the cognitive interviews.

Folder structure:

Zusammenfassung

Das Phänomen der sozial erwünschten Antwortverzerrung in Umfrageerhebungen wird in Sozialpsychologie und Sozialwissenschaften seit vielen Jahren diskutiert. Dass dieses Phänomen tatsächlich existiert, konnten in der Vergangenheit zahlreiche Forscher empirisch beweisen (z. B. Edwards, 1957; Crowne & Marlowe, 1960; Paulhus, 1984; Tourangeau, Rips, & Rasinski, 2000). Die Verzerrungen treten häufig dann auf, wenn eine Frage bzw. das interessierende Thema ‚sensitiv' ist (Lee, 1993; Tourangeau, Rips, & Rasinski, 2000) und dementsprechend einen potentiell beschämenden, belastenden oder stigmatisierenden Charakter haben (Dalton, Wimbush, & Daily, 1994: 817; Stocké, 2004). Um sozial erwünschtes Antworten bei Selbstauskünften zu vermeiden, wurden und werden weiterhin indirekte Befragungsmethoden eingesetzt und entwickelt. Diese Techniken sollen dem Befragten Anonymität und Privatheit garantieren, um dadurch ehrlichere und validere Selbstauskünfte im Hinblick auf heikle Themen zu erhalten (Warner, 1965; Lensvelt-Mulders et al., 2005; Tourangeau & Yan, 2007). Aus diesem Grund wird speziell für sensitive Themen auf eine indirekte Befragungsmethode zurückgegriffen. Hierbei handelt es sich um eine Methodik, welche unter dem Namen ‚Listenexperiment' (Kuklinski et al., 1997) oder auch ‚Item Count Technik' (Miller, 1984) bekannt ist. Mithilfe des Listenexperimentes kann auf aggregierter Ebene eine Schätzung des Anteils der Befragten, welche dem sensitiven Item zugestimmt haben, vorgenommen werden. Durch die gegebene Anonymität sollte das Listenexperiment eine höhere Zustimmungsrate (höhere Schätzung des Anteils der Befragten) zu sensitiven Themen erhalten als die direkte Befragung. Die Literatur liefert im Allgemeinen kein einheitliches Bild zur Funktionalität des Listenexperimentes, außerdem zeigen sich in wenigen veröffentlichen Studien Probleme bei der Erhebung und den Resultaten des Listenexperimentes (z.B. Biemer et al., 2005). Häufig sind die Ursachen des Scheiterns nicht ersichtlich und aus diesem Grund bildet eine Evaluation des

Listenexperimentes den Kern meiner Dissertation. Das Ziel dieser Arbeit bestand darin, geeignete Aussagen über Validität und Konsistenz des Listenexperimentes zu treffen und Faktoren/Moderatoren zu finden, welche das Scheitern des Listenexperimentes determinieren.

Die Dissertation besteht aus zwei Manuskripten, die jeweils die zentrale Forschungsfrage der Validität beinhalten. Darüber hinaus wurde der Fokus in den einzelnen Manuskripten auf aufeinander aufbauende, wie auch voneinander unabhängige Fragestellungen gelegt. In Manuskript #1 konnte auf der Basis von drei verschiedenen Studien, inklusive einer Panelanalyse mit fünf Listenexperimenten, sowie zwei unterschiedlichen Survey Modes, die Inkonsistenz des Listenexperimentes im Bereich der Vorurteilsforschung bewiesen werden. In Studie 1 (N = 229, repräsentativ, sensitives Item = Antisemitismus) lieferte das Listenexperiment zunächst theoriekonforme Ergebnisse und wies eine höhere Schätzung als die direkte Befragung auf. Studie 2, eine modifizierte Wiederholung (N = 445 (repräsentativ), konnte keine signifikante Mittelwertdifferenz innerhalb des Listenexperimentes feststellen und die direkte Befragung erzielte eine wesentlich höhere Zustimmungsrate als das Listenexperiment. Um die Validität zu prüfen, wie auch Faktoren zu finden, welche das Scheitern des Listenexperimentes bedingen, wurden in Studie 3 (N = 1.569 (nicht repräsentativ)) drei unterschiedliche Listenexperimente miteinander verglichen. Das erste Listenexperiment (sensitives Item = Antisemitismus) zeigte ähnliche Ergebnisse wie Studie 2 und galt damit erneut als gescheitert. Das zweite Listenexperiment (sensitives Item = Islamophobie) konnte eine positive, signifikante Mittelwertdifferenz vorweisen. Außerdem erzielte das Listenexperiment eine höhere Schätzung als die direkte Befragung. Mittels des dritten Listenexperimentes wurde weiterhin die Validität getestet und ein Faktor gefunden, welcher die inkonsistenten Ergebnisse erklärt. Die wesentliche Frage war, ob der Anstieg im Mittelwert auf der höheren Itemanzahl in der

Experimentalgruppe beruht. Hierfür wurden im Listenexperiment vier nicht sensitive Items gegen fünf nicht sensitive Items getestet. Um die Annahme auch valide zu testen, wurde ein auf Neutralität getestetes, nicht sensitives Item eingesetzt. Die Analyse wies eine signifikante Differenz zwischen der Bedingung mit vier und der Bedingung mit fünf nicht sensitiven Items auf. Dieses Ergebnis impliziert schwerwiegende Konsequenzen für die Validität des Listenexperimentes, da die höhere Anzahl an Items einen höheren Mittelwert provoziert und nicht allein der Inhalt des sensitiven Items für den Mittelwertanstieg verantwortlich ist. Weiterhin zeigt eine Panelanalyse zur intraindividuellen Stabilität, dass die Befragten über die Zeit konstanter antworten, wenn sich in der Kontrollbedingung lediglich vier nicht sensitive Items befinden.

Manuskript #2 war in der Lage weitere Faktoren aufzuzeigen, auf denen zum Teil die inkonsistenten Ergebnisse des Listenexperimentes beruhen. In Studie 1 wurde mittels kognitiver Interviews (N = 7) demonstriert, dass das Listenexperiment überwiegend verstanden und das sensitive Item nur zum Teil von den Befragten wahrgenommen wurde. Weiterhin zeigten die kognitiven Interviews, dass auf Grund der persönlichen Relevanz die nicht sensitiven Items für die Befragten leichter zu beantworten waren. In Studie 2 (experimentelle Onlinestudie, N = 1,878, nicht repräsentativ) wurde u. a. getestet, ob das sensitive Item das Zustimmungsverhalten zu den nicht sensitiven Items beeinflusst (Verschiebung der Itemschwierigkeit). Es zeigte sich, dass die Zustimmungsrate zu den nicht sensitiven Items steigt, wenn das sensitive Item hinzugefügt wurde. Für das Listenexperiment bedeutet dies, dass der erhöhte Mittelwert in der Experimentalgruppe, aufgrund der Verschiebung der Itemschwierigkeit durch das sensitive Item und nicht durch die vermehrte Zustimmung zum sensitiven Item entsteht. Die Hypothesen, dass die Position des sensitiven Items zur Verzerrungen in der Angabe der Anzahl der Ja-Antworten führt und zweitens, dass die Angabe der Anzahl der Items im Allgemeinen verzerrt ist, konnte in dieser Studie nicht

bestätigt werden. In Studie 3 (Replikation von Studie 2; N = 948) wurden die Hypothesen in einem leicht variierten und weiterentwickelten Design erneut getestet. Dabei konnte die erste Hypothese in einer Bedingung mit ausschließlich nicht sensitiven bestätigt werden. Die Zustimmungsrate der Befragten war in beiden nicht sensitiven Itembedingungen gleich. Desweiteren konnte bestätigt werden, dass die Angabe der Anzahl der Items, denen zugestimmt wird, im Allgemeinen verzerrt ist. In der nicht sensitiven Itembedingung ergab sich ein signifikanter Unterschied zwischen der Summe der Ja-Antworten, die aus den direkten Fragen gebildet wurde und der Angabe der Anzahl der Items. Innerhalb des Listenexperimentes bedeutet eine solche Abweichung, dass die Werte in Experimental- und Kontrollbedingung mitunter durch die Angabe der Anzahl der Items, denen zugestimmt wird, verzerrt ist.

Zusammenfassend deuten die Ergebnisse der zwei Manuskripte darauf hin, dass das Listenexperiment nicht in der Lage ist valide und konsistente Ergebnisse zu erzielen. Die Ergebnisse der vorliegenden Dissertation legen nahe, dass während der Beantwortung des Listenexperimentes Faktoren auftreten, die Verzerrungen hervorrufen und die allgemeine Funktionsweise des Listenexperimentes beeinflussen. Insgesamt wurden drei moderierende Faktoren gefunden, welche unabhängig voneinander oder gemeinsam auftreten. Die Befunde deuten darauf hin, dass das Listenexperiment in seiner momentanen Ausführung viele Probleme und unbekannte Faktoren impliziert mit denen Messfehler, sozial erwünschtes Antwortverhalten, als auch inkonsistente Ergebnisse nicht beseitigt, sondern teilweise sogar verstärkt werden.

Literatur

Biemer, P. P., Jordan, B. K., Hubbard, M., & Wright, D. (2005). A test of the item count methodology for estimating cocaine use prevalence. In Kennet, J., and J. Goefrer (Eds.), *Evaluating and improving methods used in the national survey on drug use and health (DHHS Publication No. SMA 05-4044, Methodology Series M-5)* (pp. 149-174). Rockville, MD: Dept. of Health and Human Services Administration, Office of Applied Studies.

Crowne, D., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349–354.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology*, *47*, 817–828

Edwards, A.L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

Kuklinski, J. H., Sniderman, P. M., Knight, K., Piazza, T., Tetlock, P. E., Lawrence, G. R., & Mellers, B. (1997). Racial prejudice and attitudes toward affirmative action. *American Journal of Political Science, 41*, 402–419.

Lee, R. M. (1993). *Doing research on sensitive topics.* Sage, London.

Lensvelt-Mulders, G. J. L. M., Hox, J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research. *Sociological Methods and Research, 33,* 319–348.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior.* Unpublished doctoral dissertation, George Washington University.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609.

Stocké, V. (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Rational-Choice Theorie und des Modells der Frame-Selektion. *Zeitschrift für Soziologie, 33,* 303–320.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response.* Cambridge, England: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859–883.

Warner, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60,* 63–69.

Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei allen bedanken, die mich bei der Fertigstellung dieser Arbeit begleitet und unterstützt haben.

Mein Dank gilt an erster Stelle meinem Anleiter Prof. Dr. Uli Wagner, der mich stets mit seinen Anregungen unterstützt hat und immer ein offenes Ohr für meine Anliegen und Probleme hatte. Ebenfalls herzlich danken möchte ich meinem Zweitgutachter Prof. Dr. Peter Schmidt, der mir mit wertvollen Ratschlägen und seinen Erfahrungen immer mit Rat und Tat zur Seite stand.

Darüber hinaus danke ich den Mitgliedern des Graduiertenkollegs Gruppenbezogene Menschenfeindlichkeit für die Zusammenarbeit, den Ideenaustausch und den bereichernden Tipps, die für mich immer ein sehr großer Gewinn waren. Ich danke Euch für die gemeinsame Zeit und die vielen positiven Erlebnisse, aus denen die eine oder andere Freundschaft entstanden ist. Ganz besonders danken möchte ich meinem Kollegen und langjährigen Freund Stefan Thörner für die hilfreichen Diskussionen, der konstruktiven Kritik an meiner Arbeit und für die tolle Unterstützung in allen Phasen, sowie für die lustige Zeit die wir zusammen hatten.

Mein persönlicher Dank gilt meinen Eltern, insbesondere meiner Mutter, für die unentwegte, liebevolle und tatkräftige Unterstützung, sowohl während meines Studiums als auch während der Anfertigung meiner Doktorarbeit. Danke, dass Du bedingungslos für mich da bist und mich fortwährend bestärkt hast, wenn ich an mir gezweifelt habe.

Ein ganz großes Dankeschön geht an meine Schwester Tina, auf die ich mich stets verlassen konnte und die mich immer aufs Neue ermutigt und motiviert hat. Danke für den Rückhalt, die Geduld und die unermüdliche Unterstützung, die mir viel Kraft gab.

Ich danke meinen Freunden, die mir permanent zu Seite standen, mich immer wieder aufbauten und für die nötige Abwechslung sorgten.

Erklärung des Autors

Hiermit versichere ich, dass ich meine Dissertation „Social desirability in survey research: Can the list experiment provide the truth?“ selbstständig, ohne unerlaubte Hilfe angefertigt habe und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen und Hilfen bedient habe.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen Prüfungszwecken gedient.

_____ _____

(Ort, Datum) Stefanie Gosen