

### Separating interviewer and sampling-point effects

Schnell, Rainer; Kreuter, Frauke

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

SSG Sozialwissenschaften, USB Köln

#### Empfohlene Zitierung / Suggested Citation:

Schnell, R., & Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21(3), 389-410. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-131806>

#### Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

#### Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

## Separating Interviewer and Sampling-Point Effects

*Rainer Schnell<sup>1</sup> and Frauke Kreuter<sup>2</sup>*

Data used in nationwide face-to-face surveys are almost always collected in multistage cluster samples. The relative homogeneity of the clusters selected in this way can lead to design effects at the sampling stage. Interviewers can further homogenize answers within the small geographic clusters that form the sampling points. The study presented here was designed to distinguish between interviewer effects and sampling-point effects using interpenetrated samples for conducting a nationwide survey on fear of crime. Even though one might, given the homogeneity of neighborhoods, assume that sampling-point effects would be especially strong for questions related to fear of crime in one's neighborhood, we found that, for most items, the interviewer was responsible for a greater share of the homogenizing effect than was the spatial clustering. This result can be understood if we recognize that these questions are part of a larger class of survey questions whose subject matter is either unfamiliar to the respondent or otherwise not well anchored in the mind of the respondent. These questions permit differing interpretations to be elicited by the interviewer.

*Key words:* Complex surveys; design effects; interviewer behavior; interpenetrated sample; survey sampling; interviewer variance; question characteristic.

### 1. Introduction

The majority of nationwide face-to-face surveys are conducted in multistage cluster samples, which means that respondents are clustered within small geographic areas (or sampling points). The decision to use a clustered sample (and not a simple random sample) is made for organizational and financial reasons. For example, the absence of a general population register in many countries reduces many researchers' ability to use simple random samples. Sampling in several stages, one of them at the level of small geographic clusters, allows for the selection of respondents without the aid of register data, for example with the help of random walk techniques. At the same time, interviewer travel expenses can be minimized. Clustered samples are therefore considered cost-efficient alternatives to simple random samples (Groves 1989; Behrens and Löffler 1999).

There is, however, a downside to the use of cluster samples: conventional computation of standard errors and common test procedures are based on the assumption of

<sup>1</sup> Center for Quantitative Methods and Survey Research, University of Konstanz, D 92, 78434 Konstanz, Germany. Email: rainer.schnell@uni-konstanz.de

<sup>2</sup> Joint Program in Survey Methodology, 1218 LeFrak Hall, College Park, MD 20742, U.S.A. Email: fkreuter@survey.umd.edu.

**Acknowledgments:** This project was funded by the German Academic Science Foundation (DFG) under a grant given to the first author (study number SCHN 586/2-1). We thank all project participants for their work, and the anonymous reviewers, as well as Elizabeth Bruch, Pamela Campanelli, Mick Couper, Scott Fricker, Klaus Larsen, Geert Loosveldt, Peter Lynn, Michael Mitchell and especially Elisabeth Coutts for helpful comments on this article.

independent and identically distributed observations, an assumption that is violated with clustered samples. One measure of the effect of the violation of this assumption is the so-called design effect, or the ratio of the variance of an estimator for a given sample design to the variance of an estimator for a simple random sample (Kish 1995, p. 56). The use of simple random sample formulas in the computation of significance tests and confidence intervals leads to misleading results if this design effect is greater than one, in which case standard errors will be underestimated. This problem affects not only significance tests and the computation of confidence intervals for univariate statistics, but also regression analysis, analysis of variance and goodness-of-fit tests (Biemer and Trewin 1997, pp. 608–624).

A straightforward way to explain the presence of design effects is Kish's ANOVA model (Kish 1965, p. 162), where

$$deff = 1 + roh(b - 1) \quad (1)$$

In the above equation, *roh* (*rate of homogeneity*) represents the intra-class correlation coefficient, and *b* the number of interviews conducted within the cluster. "If the variable is distributed completely at random among the clusters, then we expect a *roh* of zero, and the design effect  $[1 + roh(b - 1)] = 1$ " (Kish 1965, p. 163). *Roh* is equal to 1.0 if all elements within a cluster have the same value for a given variable. Since different questions result in different answer patterns, design effects need to be calculated on a question-by-question basis. Throughout the text we will use the design factor *deft* (the square root of *deff*) when we talk about design effects.

There is an increasing recognition of the need to adjust the confidence intervals of survey variables for the *deft*. For example, in the last several years an increasing number of statistical packages, such as SAS, SPSS or Stata, have been equipped to provide correct variance estimation for different kinds of sampling designs. The size of the *deft* reported by various researchers has shown substantial variation among surveys and variables (Kish 1995, pp. 60–61; O'Muircheartaigh and Campanelli 1998, p. 68). However, the rule of thumb employed in the survey literature is to assume a value for *deft* of about 1.4 (see, for example, Scheuch 1974, p. 39; for a discussion see Groves 1989, p. 272).

Adjusting test statistics for a design effect of  $deff = 1.96$  ( $1.4^2 = 1.96$ ) is, however, comparable to cutting the sample size almost in half. The variance of an estimator based on the data from a complex sample has the same variance the estimator would have if a simple random sample of size *n* had been adjusted by the *deft* squared:

$$n^* = n/deft^2 \quad (2)$$

A larger sample size is therefore required to obtain the same precision as one would have had with a simple random sample, i.e., to compensate for the increase in standard errors incurred through the use of a cluster sample. However, reducing the costs of conducting the survey is part of the reason for employing a cluster sample in the first place. So the question of what steps can be taken in advance to reduce design effects naturally arises, and with it the question of the sources of design effects.

### 1.1. Sources of homogeneity

Design effects result from the relative homogeneity of respondents who belong to the same cluster. At least two mechanisms are responsible for producing this relative homogeneity. First, respondents are typically more homogeneous within a cluster than in the population as a whole, whatever mode of data collection is used to obtain their survey answers. Second, the process of data collection itself can lead to an increased homogeneity within a cluster.

The first mechanism is termed spatial homogeneity. As a consequence of social processes such as self-segregation or imposed segregation according to income, age or ethnicity, respondents who live in the same geographical areas (sampling points) often are more similar to each other than respondents selected at random from the population as a whole (Lee, Forthofer, and Lorimor 1989, p. 15). There is also good reason to assume that people who live in small spatial clusters share similar attitudes because of the similar circumstances in which they live (McPherson et al. 2001). Whatever the cause of this similarity among respondents, *roh* measures the homogeneity in terms of the proportion of the total element variance that is due to group membership (Kish 1965, p. 163). If design effects are due entirely to the spatial clustering of similar people living next to each other, researchers interested in estimates with small variance may have no choice but either to give up the use of a cluster sample or to increase their sample sizes.

There is, however, a second mechanism at work in the creation of design effects, namely the data collection process. The use of a multistage cluster design involves not only the clustering of respondents, but – given the way survey research is conducted in practice – the assignment of a different interviewer to each of the respondent clusters. This interviewer may exert a further homogenizing effect within the sampling point in which he or she is active. Put somewhat differently, the interviewer could introduce an additional source of variance to the population estimators. This effect is most likely the result of the deviation of interviewer behavior from standardized procedures (Fowler and Mangione 1990, p. 28; Fowler 1991, p. 259). Interviewer technique may differ from prescribed rules in a unique way, for example with idiosyncratic interpretations of questions, incomplete reading of answer categories or incorrect probing. In addition, some interviewer characteristics can trigger other bias-producing effects such as social desirability (Schnell 1997, p. 277).

Not all questions are equally likely to be affected (Gales and Kendall 1957; Gray 1956; Hanson and Marks 1958). Researchers have tried to address the issue of questions affected by interviewers in various ways and with varying results. Results from Hyman et al. (1954), Fellegi (1964), O’Muirheartaigh (1976), and Collins and Butcher (1983), as well as those reported in Belak and Vehovar (1995), support the hypothesis that *factual items* are less vulnerable to interviewer effects than *attitude items*. This finding was not, however, replicated by Kish (1962), Groves and Magilavy (1986, p. 260) or O’Muirheartaigh and Campanelli (1998, p. 69). One possible explanation for this discrepancy is provided by Cannell (1953, discussed in Kish 1962), who suggested that interviewer effects occur if respondents are forced to answer questions about *unfamiliar topics*, especially in cases where respondents try to provide a correct or reasonable answer but they do not know which answers might be correct or reasonable. In those cases, respondents tend to use signals and help provided by the interviewer in order to gauge

the appropriateness of their answers. Similarly, Hermann (1983), using the German ALLBUS 1980, found that interviewer effects were small if the item was unimportant to the interviewer and that they were large if the item was unimportant to the respondent. Interviewer effects were found for *difficult factual items* in the 1950 U.S. Census (Kish 1962, pp. 96-97), where difficult factual items were defined as those that required the respondent to perform retrospective calculations or memory searches. Van Tilburg (1998) found large interviewer effects for questions on the respondent's personal network; these questions also belong to the group of questions that require memory search. Pickery and Loosveldt (2001) found large interviewer effects on difficult tasks such as ratings of political parties. However, data from Mangione et al. (1992) did not show larger interviewer effects for difficult items. Schnell (1997) reported that *sensitive items* that tend to evoke socially desirable answers are more likely to be affected by interviewer behavior. Bailar et al. (1977) and Fellegi (1964) showed for the National Crime Survey that items that evoke emotional reactions (for example, questions on criminal victimization) produce larger interviewer effects than questions on less emotional topics such as income or education. Gray (1956) and O'Muircheartaigh (1976) found larger interviewer effects for *open questions*, or questions for which no answer categories are provided, as those questions tend to include the interviewer in the answer process. In contrast, no support for this relationship was found subsequently by Mangione et al. (1992) or Groves and Magilavy (1986).

Given the results of past research, a well-defined classification of questions into those that tend to be vulnerable to interviewer effects and those that do not is hardly possible. One reason might be that a unidimensional classification of items is not possible (see Section 2.2 below), but instead there may be a combination of factors responsible for the effects observed. The inconsistency in the results also may be at least partly due to the fact that the design details of the surveys employed in the above studies are difficult to compare. Design details that might well have influenced the results but were not published for some of those studies included question format, question phrasing, interviewer training, number of interviews per interviewer, interviewer experience, and interviewer assignment. However, *one conclusion* seems to apply to most of those studies: interviewer effects can be reduced if the interviewer has received good training, and if the use of a standardized procedure is ensured, as emphasized by Kish (1962), Groves and Magilavy (1986), and Fowler and Mangione (1990).

The strongest interviewer effects appear with interview procedures that are not completely standardized (Hyman et al. 1954): if the interviewer is, for example, given latitude regarding decisions on the use of neutral answer categories (Collins 1980), the coding of answers (Rustemeyer 1977) or the phrasing of individual probing questions (Mangione et al. 1992). Such findings lead to the following hypotheses:

- (1) For the majority of questions, active interviewer involvement is responsible for most of the observed interviewer effects. The interviewer effects most often described in the literature refer to a different process. These interviewer effects result from the reaction of the respondent to a perceived relationship between obvious interviewer characteristics (for example, race, age, sex or regional dialect) and the content of the question. No active interviewer involvement is required to

produce such effects. Active interviewer involvement, on the other hand, is any deviation from prescribed interview protocol, such as reformulating the question or aiding in its interpretation. Interviewers may involve themselves in anticipation of an adverse reaction by the respondent or the respondent may prompt them for help.

- (2) Active interviewer involvement in question clarification is more likely when the respondent requests additional interpretation and explanation of a question.
- (3) Respondents are likely to ask for clarification when
  - (a) They do not understand the question wording,
  - (b) They are asked to answer a factual question on a topic about which they know nothing,
  - (c) Ambiguous terms are used in a factual question or
  - (d) The item is an attitude question and the respondents' attitudes are not salient.

As a consequence, one would assume that certain interviewer effects occur not just with a specific item type (attitude vs. factual, for example) but with a certain combination of item types or target populations and item types (for example, some item types may be problematic for certain subpopulations but not for others). The size of an interviewer effect therefore depends on interviewer characteristics, target population, question topic, and question type. Here we will concentrate on the last two variables. Questions that are difficult for respondents, or are perceived as such by interviewers, will yield larger interviewer effects. To be more precise:

- (4) Larger interviewer effects are to be expected if either the topic of the question is unfamiliar or the answer is not well anchored in the mind of the respondent.

Such questions leave more room for differing interpretations to be elicited by the interviewer (Martin 1983; Fowler and Mangione 1990; Mangione et al. 1992; Schnell 1997; Kreuter 2002).

### *1.2. Interviewer and sampling-point effects*

Even though the potential for interviewers to amplify measurement errors was first addressed many years ago (Rice 1929) and there have been several warnings in recent decades about interviewer effects (Kish 1962; Hedges 1980; Hageaars and Heinen 1982; O'Muircheartaigh and Campanelli 1998), little has been done to reduce interviewer effects in the practice of face-to-face surveys. This neglect of the role of the interviewer in the overall design effect in clustered (face-to-face) surveys might be due to the fact that the combination of multistage cluster designs commonly used in survey practice and the interviewer assignment practices routinely employed do not allow for the determination of the relative effect sizes of the two sources.

Interviewer effects combine with the effect of spatial homogenization to produce the design effects observed for a multistage cluster sample. Given the way interviewer assignment functions in practice, these two contributions to the design effect are difficult to separate: interviewers are usually assigned to one sampling point, and they seldom work in another sampling point unless they are employed in a large city in which interviewers

can easily be sent out to help each other carry out the required number of interviews. Since interviewers generally canvas only one point and a given point is covered only by one interviewer, determining how much of the observed variance of a given estimator is due to spatial proximity or the interviewer is most often impossible (Hoag and Allerbeck 1981, p. 414; Groves 1989, p. 271). Therefore, little is known about the separate contributions of the interviewer and the geographic area to the total design effect. In order to estimate the interviewer effect in geographically clustered face-to-face surveys, one must use interpenetrated samples (Bailar 1983, pp. 198–199). A simple version of this design would ensure that the addresses within each sampling point are randomly assigned to several interviewers working independently and that every interviewer is assigned to one single sampling point (Biemer and Stokes 1985, p. 159).

Only a few studies have been designed – with different kinds of interpenetrated samples – explicitly to estimate the relative impact of interviewer and sampling point on the design effect (Hansen et al. 1961; Bailar et al. 1977; Bailey et al. 1978; Collins and Butcher 1983; Davis and Scott 1995; O’Muircheartaigh and Campanelli 1998). These have attempted to separate the effects of interviewer and sampling point on the variance in the answers to individual survey questions ( $i$ ) observed for a given cluster ( $\sigma_{i\text{cluster}}^2$ ):

$$\sigma_{i\text{cluster}}^2 = \sigma_{i\text{Sampling-Point}}^2 + \sigma_{i\text{Interviewer}}^2 \quad (3)$$

All of the above studies showed that at least part of the cluster variance (in answers to specific questions) within a given sampling point was due to the interviewer ( $\sigma_{i\text{Interviewer}}^2$ ). Knowing that interviewers play a role in producing design effects, one might ask how large this role is in survey practice.

More information on the magnitude of interviewer effects is available from the analysis of data from a series of telephone surveys (e.g., Kish 1962; Gray 1956; Hanson and Marks 1958; Tucker 1983; Groves and Magilavy 1986). Telephone surveys are employed for the study of interviewer effects since potential respondents from all sampling points – if there are even any defined – can easily be assigned to interviewers randomly. In the analysis of telephone survey data, respondents are clustered by interviewer.  $Roh$ , as it applies to interviewer effects, is the proportion of total element variance that is due to the interviewer. The average value of  $roh$  for these surveys is 0.01, and according to Groves (1989, p. 318), values above 0.1 are seldom observed. However, a small value of  $roh$  in telephone surveys should not lead to the impression that interviewer effects do not lead to design effects. The average interviewer workload varies a great deal, and it is not unusual to find an interviewer workload in a telephone survey of between 60 and 70 interviews per interviewer (Tucker 1983; Groves and Magilavy 1986). Again using Equation (1) to compute the design effect, an interviewer workload of 70 interviews would lead to  $deft^2 = 1 + 0.01 \cdot (70 - 1) = 1.69$ . Whatever the actual size of the interviewer effect for telephone interviews is, it is questionable whether the interviewer effects found with telephone surveys can be generalized to the face-to-face situation, where cost-efficiency forces survey institutes to employ cluster samples. The study presented here therefore represents an extension of the above studies in its attempt to separate the two sources of design effects for nationwide face-to-face interviews.

## 2. Data and Method

The data presented are part of the design effect study (DEFECT), designed to study sampling errors and nonsampling errors in complex surveys. The DEFECT study is the first nationwide sample to be carried out in Germany with an interpenetrated sampling design.<sup>3</sup> The same questionnaire was used to conduct five independent surveys in 160 sampling points. Four of the five surveys were conducted by professional survey institutes that are highly regarded for their good practices and reliability, while the fifth (the mail survey) was conducted by the DEFECT group itself. This design means that in each of these 160 sampling points, two face-to-face surveys with random sample selection, a face-to-face survey with quota selection, a CATI, and a mail survey (both with random selection) were carried out concurrently (the use of several survey modes and sampling procedures has to do with the fact that this study was conducted as part of a methodological study with a much larger scope). A schematic overview is given in Table 1. For details on the study, see Schnell and Kreuter (2000) and the project homepage ([http://www.uni-konstanz.de/FuF/Verwiss/Schnell/DEFECT\\_home.html](http://www.uni-konstanz.de/FuF/Verwiss/Schnell/DEFECT_home.html)).

For each of the three face-to-face surveys, only one interviewer was active in each sampling point and no interviewer worked in more than one sampling point. This means that overall three interviewers worked within each sampling point independently of each other. The use of this design was intended to permit the statistical separation of interviewer and sampling-point effects, while adhering to the survey institutes' usual procedures. A cross-classified design, one in which one interviewer was assigned to more than one sampling point, would have been not only unusual but difficult to implement for a face-to-face survey. A telephone survey can be easily adapted to multi-sampling-point interviewer assignment, but the use of a cross-classified design in a face-to-face survey would be substantially more challenging. Such a design would have been reasonable only if the sampling points to which an interviewer was assigned had a substantial geographic separation, since assigning the same interviewer to neighboring points would not have allowed for a valid separation of sampling-point and interviewer effects. Assigning one interviewer to sampling points in different regions would, however, have involved considerably higher travel and administrative expenses. It should be mentioned here that the participating survey institutes incurred already a financial loss even in working with the sampling design actually employed.

### 2.1. Sampling

The addresses of households to be contacted in these surveys were selected in a multistage procedure. In the first stage, 160 PSUs (sampling points) were randomly selected from a nationwide register of election districts. In the second stage, an address-random procedure was used to select households within the 160 election districts. The actual household selection was carried out by eight members of the research group, who personally visited each of the sampling points. Starting from a randomly selected address in each sampling

<sup>3</sup> To our knowledge, the only other nationwide survey that employed a design that may be considered an interpenetrating sample design was the 2nd wave of the British Household Panel 1991 (O'Muircheartaigh and Campanelli 1998). Their study design appears to differ from ours in the number of PSUs and their selection.



*Table 1. Design of the DEFECT Surveys*

	Sample selection	Mode	Mode for nonresponse study
Survey I	Random	Face-to-Face	CATI
Survey II	Random	Face-to-Face	CATI
Survey III	Quota	Face-to-Face	
Survey IV	Random	CATI	CATI
Survey V	Random	Mail	CATI

point, the group members noted the address of every third household along the random-walk route until they had gathered 110 addresses. Of these 110 addresses, the first 64 (16\*4) in each point were randomly assigned to the four random surveys mentioned above using a shuffle procedure, which meant that a total of 2,560 addresses were initially sent to the survey institutes. Shortly afterwards, four additional addresses per sampling point were given to the institutes for replacement (for example to replace respondents who had either died or moved away since the address collection had been conducted). In some sampling points, even this number of addresses was not enough to permit a total of at least six realized interviews in all sampling points. In such cases, the institute received additional addresses for the points. However, the interviewers were committed to making at least four contact attempts at each of the original addresses before this fallback sample could be used. The quota survey did not employ the addresses collected during the random walk.

The addresses received by the institutes were therefore household addresses, but the goal was an individual sample of the target population. That target population was defined as all German-speaking inhabitants of those households who were 18 years or older, one of whom was to be randomly selected for survey participation. The institutes themselves were responsible for this last sampling stage. In the face-to-face random surveys, the potential respondent was selected by the interviewer using a variant of a Kish grid. For the other two random surveys (CATI and mail), the potential respondent was selected using the last-birthday procedure.

In keeping with the desire to employ an interpenetrated sampling design, the general policy of this study was that only one interviewer should contact the households in a given point on behalf of a certain institute. In practice, this requirement could not be met by the institutes in every single sampling point. For example, one of the interviewers refused to return to a certain sampling point after his hubcaps were stolen there during his first visit. Given the relevance of such incidents to fear of crime (the topic of the survey), sending a replacement interviewer was, of course, imperative. The institutes were therefore allowed to assign a second interviewer to a point, but not in more than 10 percent of the sampling points. The survey institutes agreed to the stipulation that as soon as a second interviewer was assigned to a point, the first interviewer had to stop working there. This requirement, along with the procedures dictated by the original agreement, meant that at any given time there was only one interviewer from each institute working in each sampling point.

The main data sources for the following analysis are data from the two face-to-face surveys with random selection. Overall, 1,345 and 1,326 interviews were conducted; representing 39.3 and 41.5 percent of the eligible respondents (see Table 2). The mail survey had a response rate of 48.7, which is 1,152 responses out of the 2,364 eligible

Table 2. Response rates for the face-to-face surveys

	Institute I		Institute II	
	<i>n</i>	%	<i>n</i>	%
Addresses delivered to the institutes	4,889	100.0	3,868	100.0
Unused addresses	1,231	25.2	445	11.5
Gross sample	3,658	100.0	3,423	100.0
Ineligible households	232	6.4	230	6.7
Gross sample of eligible households	3,426	100.0	3,193	100.0
Unable to reach household	582	17.0	735	23.0
Unable to reach respondent/respondent unable to do interview	88	2.6	79	2.5
Appointment scheduled after end of field period	–	–	11	0.3
Household refusal	1,161	33.9	537	16.8
Respondent refusal	241	7.0	423	13.2
Interview incomplete	2	0.1	7	0.2
Interview invalid	4	0.1	71	2.2
Other	3	0.1	4	0.1
Successful interview	1,345	39.3	1,326	41.5

respondents. The marginal distributions of demographics and responses are similar in the two face-to-face surveys and the mail survey. A detailed analysis of each survey's nonresponse study will lead to further insight (Rässler and Schnell 2004).

## 2.2. Survey content and item classification

The survey itself was on fear of crime. This topic was chosen for this methodological study for several reasons. First, it was intended to boost participation since crime is a matter of at least some concern to a wide variety of people. Second, several kinds of questions can be asked on fear of crime (factual, attitudinal, sensitive, and so forth), the use of which allows for a comparison of design effects for different kinds of items. In constructing the questionnaire, we used both the well-established indicators typically used in fear of crime surveys and a set of items we developed ourselves (Kreuter 2002). All questions went through three pretest phases. In the first phase, we conducted a series of cognitive pretests to ensure that respondents understood the question wording, the use of the answer scales and the like. Unclear questions were rephrased and tested again. In the second phase, we conducted several additional rounds of pretests; in each of these rounds, five to 20 respondents filled out the complete paper-and-pencil version of the questionnaire. The questionnaire was evaluated and improved on the basis of both the behavioral coding conducted during the pretest and the answers to a number of probing questions asked afterwards. In the third phase, the finalized questionnaire was tested in a large-scale telephone survey for question flow, question-order effects, and completion time. A detailed documentation of the questionnaire development was created using a newly developed software tool called QDDS (questionnaire development documentation system, see Schnell and Kreuter (2001)). The application of this tool to the DEFECT questionnaire

includes all versions of the questionnaire and can be viewed (in German) at <http://esem.bsz-bw.de/sicher>.

The final questionnaire for the respondents contained 71 questions reflecting many different item types, with binary, categorical, and continuous answer scales. Given the fact that each question could contain multiple items, the questionnaire contained 135 items, if all filters are counted. Among them were questions on fear of crime itself, subjective victimization risk, prior primary and secondary victimization experience, and coping measures taken by the respondent, along with other questions on the respondent's housing situation, health, household composition, and standard demographic characteristics. The face-to-face interviewers in the DEFECT study were asked to answer additional questions about each respondent. Those questions are not used in the analysis presented here.

Each of these 135 items was classified by four members of the research project into 16 item types derived from the classification scheme suggested by Mangione, Fowler, and Louis (1992), which includes four dimensions: difficult/easy, sensitive/nonsensitive, factual/nonfactual and open/closed. Each item was rated on all four dimensions. If the dimensions are considered individually, 84 items were classified as factual, 67 as sensitive, 40 as difficult, and 14 as open. Items were classified as difficult if the respondent was asked to produce an answer that might tax his or her recall abilities (for example, the number of visits to the doctor in the last three years) or if the respondent was confronted with a complicated issue about which he or she might have never thought before (for example, how often he or she had worried about victimization in the last two weeks). Items were classified as sensitive if a certain response could be seen as socially desirable (for example, less fear for a fear-of-crime question) or if the question was unpleasant to answer (for example, previous victimization experience). Items were classified as factual if a check of the answers was, at least in theory, possible (for example, the presence or absence of home-security devices). Items were classified as open if no response options were read out loud to the respondent or if a numerical estimate was requested (for example, the subjective victimization probability in percent terms).

### 2.3. Method

The dependent variables in the analyses presented here are the square root of the design effects (*defts*) and ratios of variances. *Defts* are employed as a measure of the effect of using a complex sample design. For the current analysis, we used several techniques to compute the design effects: Taylor linearization, bootstrap, random groups, and jackknife procedures (Kish and Frankel 1974; Shao 1996). The results reported here were produced using only the Taylor linearization, since the differences among the results produced by these techniques were negligible and, according to Kish (1995, p. 57), *defts* should be viewed as rough measures for large effects. For binary items, Taylor linearization provides a good approximation for design effects if they are not heavily skewed (Goldstein et al. 2000).

In a cross-classified design, interviewer and sampling-point effects could be compared by first computing design effects using the interviewer as the cluster-defining variable and then comparing those to design effects estimated using the sampling points as the cluster-defining variable. However, the DEFECT study design was not conceived of as a cross-classified model, but rather as a nested hierarchical model. Therefore, a three-level model

was used to estimate the relative effect of interviewer and sampling point. The respondents form the lowest hierarchical level, the interviewers the second level, and the sampling points the highest level in this model. The assumption is that the mean of the variable of interest can be estimated from an interviewer specific effect  $k_{pi}$  within a sampling point and a random error  $e_{pir}$  for the respondents surveyed by each interviewer within every sampling point (see Equation (4)). The value for  $k_{pi}$  is estimated from a constant  $\mu$  and two random effects:  $a_p$  for the sampling points and  $b_{pi}$  for the interviewer, where  $p$  can take values from 1 to  $P$  (the total number of sampling points),  $i$  values from 1 to  $I$  (the total number of interviewers within any given sampling point, here a maximum of two), and  $r$  values from 1 to  $R$  (the total number of respondents interviewed by the same interviewer within any given sampling point), and where  $a$ ,  $b$ , and  $e$  are independent of one another. The distribution of the random effects is assumed to be normal. In the interest of clarity, we omit the item subscript in the following equations:

$$y_{pir} = k_{pi} + e_{pir} \quad (4)$$

$$k_{pi} = \mu + a_p + b_{pi} \quad (5)$$

The model for each binary item is listed below, where  $u$  represents the random effect for the sampling points and  $v$  represents the random effect for the interviewers, and  $u$  and  $v$  are independent and  $y$  is conditionally independent given  $u$  and  $v$ :

$$P(y_{pir} = 1 | \pi_{pi}) = \pi_{pi} \quad (6)$$

$$\text{logit}(\pi_{pi}) = \beta + u_p + v_{pi} \quad (7)$$

A ratio that we will call  $R_I$  was computed from the resulting variance components ( $a$  and  $b$  in the linear case; and  $u$  and  $v$  for the binary case).  $R_I$  is the ratio of the variance component that is due to the interviewer ( $\sigma_I^2$ ) over the sum of the variance components due to the interviewer and to the sampling point ( $\sigma_p^2 + \sigma_I^2$ ).

$$R_I = \sigma_I^2 / (\sigma_p^2 + \sigma_I^2) \quad (8)$$

Computing  $R_I$  in this way permits the comparison of the proportion of the variance due to the interviewer across all items.

In order to prepare the data for the analysis, several adjustments were made to the data set. First, answers were transformed for several variables. Nonordinal categorical items with more than two categories were split up into several indicator variables and treated in the same way as all other binary variables. Ordinal attitude scales were treated as continuous. Similar procedures were used by both Davis and Scott (1995, p. 42) and O'Muircheartaigh and Campanelli (1998, p. 65). Second, in order to avoid numerical estimation problems, highly skewed variables were excluded from the analysis; we considered binary items to be highly skewed if more than 90 percent of the answers to those items were in one of the two categories.

Of the original 132 variables, 14 binary variables were excluded because of skewness. Ten nonordinal categorical variables were transformed into 86 binary indicator items, of which 58 were dropped since only a few respondents picked those answer categories, which made them highly skewed as well. Aside from the highly skewed binary items and

indicators, twelve additional variables were excluded as well – two highly skewed noncategorical items, one whose answer categories unintentionally differed between the two face-to-face surveys, others because their answers were constrained by design (for example the “number of inhabitants in the sampling point”) or because the answers were determined entirely by the interviewer contact strategy (for example, “day of the week on which the respondent was interviewed”). Overall, 123 items were left for the analysis.

Those sampling points in which the number of interviews conducted by either institute was fewer than six were excluded to ensure more or less equal cluster size. That left 132 sampling points, in which 264 interviewers had interviewed a total of 2,280 respondents for the analysis. Finally, only those surveys that exclusively employed the interpenetrated random sample design described above were analyzed here, which means that data from the quota and CATI surveys were excluded. For comparative purposes, data from the mail survey were analyzed as well, since any design effects found in the DEFECT mail survey can clearly be attributed to the homogenizing effect of spatial clustering. This is so because the DEFECT mail survey was conducted in the same sampling points and cannot, by definition, have any interviewer effects.

Three estimation procedures were used to decompose the variance of the items from the combined face-to-face surveys into sampling point and interviewer components. First, iterative generalized least-squares algorithms were employed using MLwiN. Second, for binary variables Marginal Quasilielihood (1-MQL) and Penalized Quasilielihood (2-PQL), as implemented in MLwiN (Rasbash et al. 2000), were used. Third, since MQL and PQL tend to underestimate random effects in the case of smaller clusters (Rodríguez and Goldman 1995), numerical integration using adaptive Gaussian quadrature (AGQ) was used as well. AGQ is implemented in the *gllamm* module that is available for Stata (Rabe-Hesketh et al. 2002). For the items used in this analysis, the results obtained using the various techniques were very similar. The absolute difference in  $R_I$  between PQL and AGQ was below 0.15 for all but eight items, of which four did not achieve convergence with MQL/PQL. There were larger differences for the other four items; these four items were excluded from further analysis. For the remaining items, the average difference between PQL and AGQ was 0.035. A large difference also was obtained for one of the continuous items depending on whether the *gllamm* or MLwiN estimation was used; the average difference for all other items was 0.03. The results presented below were obtained using the adaptive Gaussian quadrature for all the 118 remaining items. The code will be provided upon request.

### 3. Results

Regarding face-to-face surveys in the DEFECT study, the mean design effect for the 118 items in each survey is 1.39, with a standard deviation of 0.27 (and median values of 1.34 for one face-to-face survey and 1.36 for the other). The mean design effect computed for the pooled sample of the two face-to-face surveys using sampling points as primary sampling units is 1.48 (median 1.43). These values correspond well to the above-mentioned rule of thumb.

Most of the design effects found with these data are attributable to the interviewers. After splitting up the variances obtained using a three-level hierarchical model, the

relative share of the total variance due to the interviewer and sampling point combined varied around a mean of 0.77, and a median of 0.82. This means that, for those 118 items used in the DEFECT study on fear of crime, most items showed a larger relative proportion of interviewer-induced variance than geographic-clustering-induced variance. Figure 1 depicts the design effects for all of the items used in the DEFECT study plotted against the interviewer proportion of the total variance (the variance due to both interviewer and the sampling point).<sup>4</sup>

Given the usual (although mostly implicit) assumption of the importance of geographic clustering, this result is surprising. The plausibility of this result is, however, supported by evidence from two comparisons. First, we examine whether the estimated fraction of interviewer variance is higher for those items for which previous work would predict a greater susceptibility to interviewer effects. Second, we compare the size of the design effects to those estimated using the mail survey, since the DEFECT mail survey was conducted in the same sampling points and cannot, by definition, have any interviewer effects.

### 3.1. Assumed mechanisms for interviewer effects

The first line of evidence concerns the theoretical plausibility of the size of the interviewer effects and design effects depicted in Figure 1. Items in Figure 1 that have a very large overall design effect (we consider  $deft > 2$  to be very large) are those that are closely related to the sampling point. These include items such as those on incivility (for example, the presence of abandoned houses or graffiti in the respondent's neighborhood), home ownership status, and distance to the closest train station.

Although these items share high overall design effects, the proportion of the total cluster variance that is due to the interviewer varies among the items. The large design effects observed for the home ownership indicator variables ( $deft \sim 2.5$ ) are almost entirely attributable to the sampling point ( $R_I \sim 0.13$ ). Incivility items, however, show a larger proportion of interviewer variance than sampling-point variance ( $0.49 < R_I < 0.58$ ), which is also the case for the item on the distance to the closest train station ( $R_I = 0.67$ ). Among those items that show a very low overall design effect are questions on the number of household members or the respondent's victimization experience within the last twelve months. The (low) design effect for the former has almost no interviewer contribution. The small design effect for the latter, however, is almost entirely attributable to the interviewer.

The high proportion of cluster variance due to interviewers might, at a first glance, seem surprising for an item like "distance to the closest train station" since this item is a factual one that should theoretically not be susceptible to interviewer effects. It was, however, also classified as a difficult item that might require help from the interviewer, especially since no closed answer categories were provided. The gross effects of different item properties can be seen in Figure 2, which depicts  $R_I$  for the four item types.

We consider only those items for which a sizeable overall cluster-level design effect was found, since it would not be meaningful to discuss the proportion of the total cluster

<sup>4</sup> The careful reader might notice some items with design effects below one. This finding is, although empirically rare, theoretically possible (Kish 1965, p. 163). The substantive explanation of such cases is not, however, straight-forward.

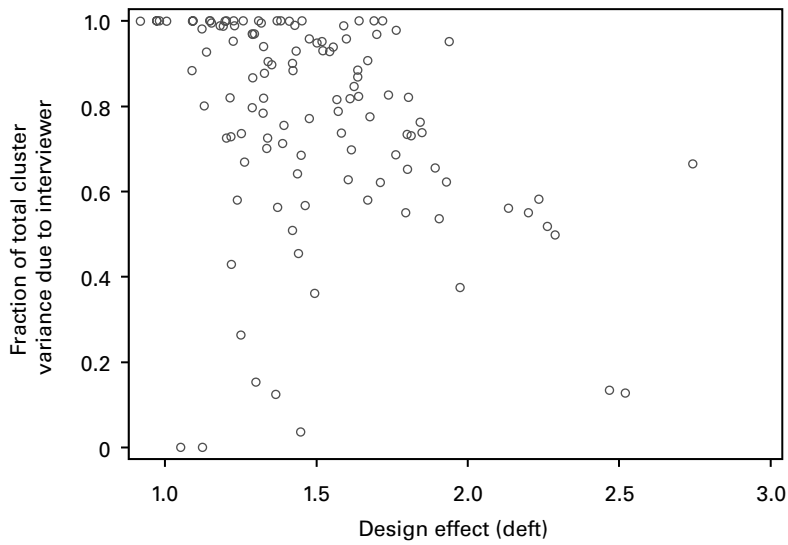


Fig. 1. Fraction of interviewer variance and overall design effect (displayed as *deft*) for 118 items

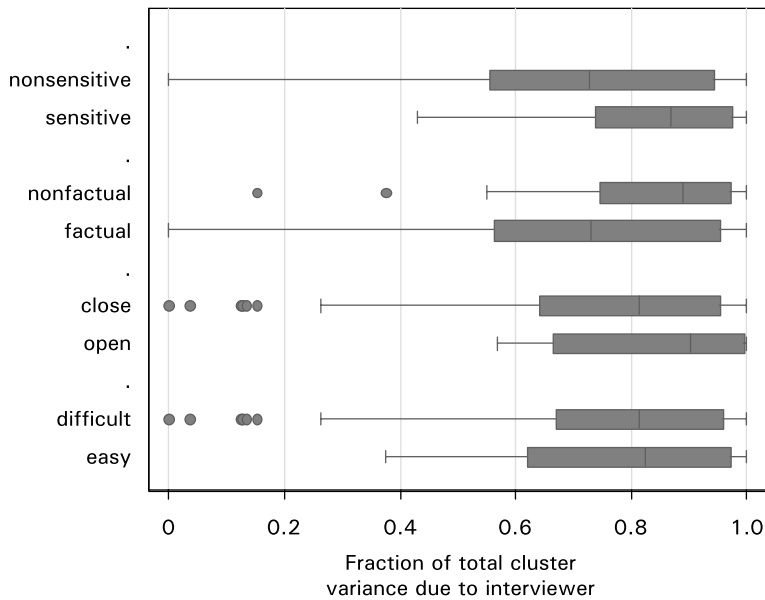


Fig. 2. Box plots of interviewer variance as a fraction of variance due to both interviewers and sampling points grouped by item type

variance due to the interviewer in cases where the total variance is very small. We therefore excluded all items for which the overall *deft* was smaller than 1.1.

In keeping with our hypotheses, we would expect a higher proportion of the total cluster variance to be due to interviewer variance for nonfactual items than for factual items (a), and for items that were classified as sensitive than for those that were classified as

nonsensitive (b). We also would expect relatively larger interviewer effects for difficult items than for easy items (c) and for items without closed answer categories (d).

Even if tests of only the first two hypotheses (a and b) yielded significant differences, the data provide support for three of the four hypotheses (see the grouped box plots in Figure 2, and the test results: *t*-tests (one-sided):  $p < 0.001$ ,  $< 0.003$ , 0.13, 0.25; Kruskal-Wallis:  $p = 0.02$ , 0.01, 0.27, 0.8).

In summary:

- Sensitive items produce higher interviewer effects than nonsensitive items (mean  $R_I$  is 0.84 vs 0.69).
- Nonfactual items produce higher interviewer effects than factual items (mean  $R_I$  is 0.83 vs 0.71).
- Open questions produce higher interviewer effects than closed questions (mean  $R_I$  is 0.84 vs 0.75).
- No differences in interviewer effects were found for difficult and easy items ( $R_I$  is 0.78 vs 0.75).

The number of items was not sufficient for an ANOVA with all the interaction effects of all four of the item properties (sensitivity, factuality, openness, and difficulty). Nevertheless, it should be noted that the difference between factual and nonfactual items was larger among the nonsensitive questions (0.65 vs 0.83) than among the sensitive ones (0.84 vs 0.84). Interaction effects among the various item characteristics therefore may reasonably be expected for a similar analysis of interviewer effects that employs a larger number of items.

Despite the fact that possible interaction effects were not examined, a comparison of the results obtained here with findings from previous work on interviewer effects could be instructive. We therefore defined a simple additive index of possibly harmful item properties. This index is intended to capture the amount of a question's slackness during the interpretation process. It reflects the likelihood that the question provokes either (a) unsolicited cues from the interviewer due to his or her perception of its problematic nature, (b) requests for the interviewer's clarification or help by the respondent or (c) use by the respondent of the interviewer's nonverbal feedback for the determination of an appropriate answer.

An item need not have all four properties to be susceptible to interviewer effects. However, a question with more than two possibly harmful properties may be assumed to be more susceptible to interviewer effects than a question with only one or two such properties. The lowest interviewer effects therefore should be seen for items with no potentially harmful properties, i.e., for those items that are factual, nonsensitive, closed, and easy to answer. Naturally, the number of possibly harmful item properties varies between 0 and 4. Since only four items had all four of these properties, we pooled items with three or four of the properties into one category.

The fraction of the total cluster variance that is due to interviewer variance  $R_I$  varies remarkably across the item groups (see Figure 3). Box plots and an ANOVA show significant differences among the various  $R_I$  calculated for the four item groups in the pooled DEFECT face-to-face survey (see Table 3). There appears to be a steady increase



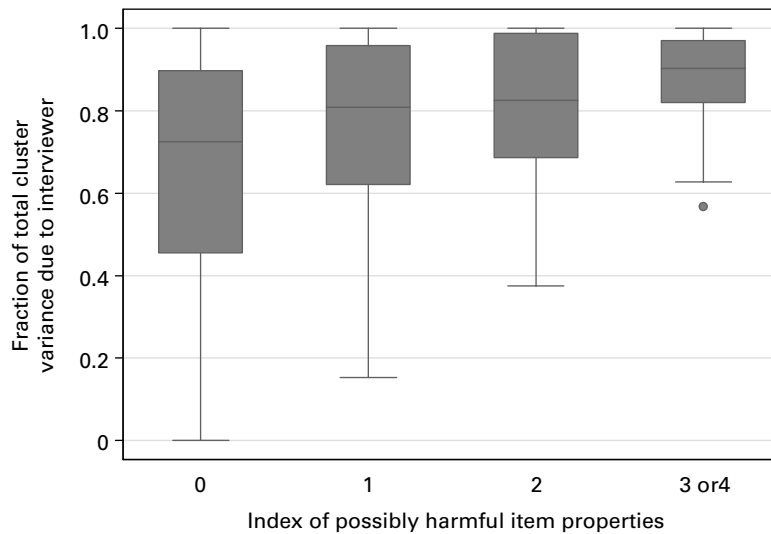


Fig. 3. Box plots of proportion of interviewer variance grouped by number of possibly harmful item properties

Table 3.  $R_1$  for items grouped according to presence of possibly harmful item properties

Number of possibly harmful item properties	DEFECT Face-to-face surveys	
	Mean $R$	$n$
0	0.63	29
1	0.78	34
2	0.82	27
3	0.87	18

ANOVA results  
 $F = 4.95$  (df = 3)  
 $P = 0.003$ ,  $r^2 = 0.1$

in the size of interviewer effects with the number of possibly harmful item properties. Cuzick's (1985) nonparametric trend test yields a  $z$ -value of 2.81.

### 3.2. Comparison of the face-to-face-surveys with a mail survey

We can exploit the design of the DEFECT study for yet another check of the plausibility of the findings. The DEFECT mail survey was conducted at the same time and in the same sampling points as the face-to-face survey, using the same questionnaire. This fact allows us to make direct comparisons between the size of design effects found in the DEFECT face-to-face surveys and those found in the DEFECT mail survey. Nationwide clustered mail surveys of a general-population sample are rather rare since usually there are no cost advantages to employing a cluster design for a mail survey. However, data from a clustered-sample mail survey can be used to gain an estimate of the size of the design effects produced by spatial clustering. Design effects found in the mail survey are, by definition, solely sampling-point effects.

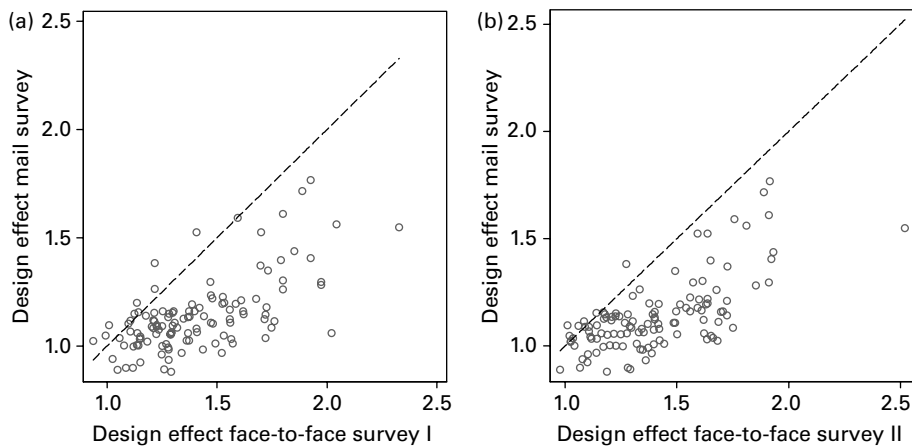


Fig. 4. Design effects for all items in the mail survey vs. effects for the same items in each of the two face-to-face surveys

The *defts* found for the mail survey have a mean of 1.1 (median 1.1) and a standard deviation of 0.17. These values are considerably lower than the ones found for both the individual face-to-face surveys and the pooled data for both face-to-face surveys. Both Figure 4a and Figure 4b display the design effects estimated for 117 items in the DEFECT mail survey against the design effects estimated for the same set of items using each of the two DEFECT face-to-face surveys. The same subset of points and items was used for the analysis of the mail survey as for the analysis of the face-to-face surveys. There were, however, too few observations for one of the mail survey indicator items on occupational status for it to be used in the analysis, meaning that the comparison uses 117 of the 118 face-to-face items.

In both comparisons, only very few items fall above the identity line, i.e., produce a higher design effect in the mail survey than in the face-to-face survey, and those that do are those that had small design effects in the first place. The highest *deft* found for the mail survey is 1.77. Only nine mail-survey items produced a *deft* larger than 1.4, which was the average design effect size found for the face-to-face surveys.

Of those nine items, three are related to the presence of incivility factors in the neighborhood (for example, the presence of graffiti), two are the home ownership indicators, one refers to the presence of motion detectors around the house, two are related to questions on the closest train station, and only one item is attitudinal. The attitudinal item requires the respondents to compare the safety of their city with that of other cities. All of these items are strongly related to the geographic clustering of the sampling point, and design effects would therefore be expected.

For those nine items, the design effects computed for all three surveys are displayed in Table 4. The data reveal that, for both face-to-face surveys, eight of the nine items have design effects that are as high or higher than the design effects computed for those items in the mail survey. Given this result, and given the difference between the average *deft* observed for the mail and face-to-face surveys over the entire item set (i.e., 1.1 versus 1.4), we can infer the influence in the face-to-face surveys of a homogenizing force acting in addition to spatial cluster effects.

Table 4. Design effects found for the mail survey versus both face-to-face surveys

	Deft mail survey	Deft face-to-face I	Deft face-to-face II
Signs of loitering	1.44	1.85	1.93
Safety in home city compared to other cities	1.52	1.4	1.59
How often at train station	1.53	1.7	1.64
Distance to train station	1.55	2.33	2.52
Graffiti visible in neighborhood	1.57	2.04	1.81
Home security: Motion detector	1.59	1.6	1.76
Empty or abandoned houses	1.61	1.8	1.91
Home ownership status	1.72	1.89	1.89
Is the renter the main tenant?	1.77	1.93	1.92

Another line of argument would hold that the cluster effects observed for mail-survey items should be independent of the item dimensions suggested above, since the effect of the item type is assumed to be exerted through the interviewer. If this logic holds, there should be no difference in the average design effect for items differentially classified according to the number of potentially harmful properties they have. Since there are no interviewers in a mail survey, a computation of  $R_1$  has no meaning here. The design effect is a pure clustering effect. The same computation with the data from the face-to-face surveys would confound interviewer and cluster effects. Therefore, we do not compare *defts* of mail and face-to-face surveys here directly.

We ran an ANOVA for the mail survey data. Again, the analysis was done only for items that showed a considerable design effect ( $deft \geq 1.1$ ), which meant that data for 60 mail survey items were analyzed. As predicted, we found no effect for item type (see Table 5).

The effect of possibly harmful item properties on design effects therefore seems to be specific to the data collection mode: the size of the effect depends on the presence and behavior of an interviewer. However, since mail surveys of the general population are sometimes considered to be biased toward the more educated, there still may be other reasons for using a face-to-face survey instead of a mail survey.

Table 5. Deft for items grouped according to presence of possibly harmful item properties in the mail survey

Number of possibly harmful item properties	DEFECT mail survey	
	Mean deft	<i>n</i>
0	1.29	14
1	1.26	26
2	1.22	18
3–4	1.14	4

ANOVA results:  
 $F = 1.16$  (df = 3)  
 $p = .334$ ,  $r^2 = 0.06$

#### 4. Conclusions

Even though one might reasonably assume that design effects for items in a survey on fear of crime in one's neighborhood would be due mostly to the homogeneity of those neighborhoods, our analysis showed that, for most items, the interviewer was responsible for a larger proportion of the homogenizing effect than was spatial clustering. This somewhat surprising result can be better understood by looking at the influence of item type on interviewer effects. There were systematic influences on the size of the interviewer effects found, since different item types subject the respondent to different cognitive or emotional burdens. If an item is sensitive, nonfactual, difficult or open, the question leaves more room for interpretation and is more difficult to answer, meaning that the respondent is more likely to rely on explicit or implicit help from the interviewer. These items therefore produce larger interviewer effects, especially if two or more of these item properties are present. Support for the large interviewer effects we found was provided by comparing the design effects found for a mail survey to those found for two face-to-face surveys. The design effects found for the clustered-sample mail survey are considerably lower than those found for the face-to-face surveys.

Several practical conclusions can be drawn from these results. First, there is reason to argue that the design effects commonly observed in face-to-face surveys are not unavoidable. Rewriting particularly susceptible questions to reduce the interviewer's potential influence on the respondent's answers might be a worthwhile improvement, as might be reducing interviewer workload. Both measures would reduce design effects. Second, interviewer and sampling-point identifiers should be included as variables in every data set, since their inclusion would enable the computation of corrected standard errors using two different kinds of clusters. This suggestion stands in contrast to current practice, in which corrected standard errors are not estimated using the interviewer identifier as the cluster-defining variable but by using the PSU as cluster-defining variable. Third, given the large effects found for interviewers, the computation of corrected standard errors might also be necessary for the analysis of data from telephone surveys. This procedure does not appear to be a part of current practice. However, further research is needed to determine to what extent interviewer effects in telephone surveys are comparable to interviewer effects in face-to-face surveys. Those comparisons are possible with the DEFECT data and such analysis is forthcoming. Preliminary datasets are available on request.

#### 5. References

- Bailar, B.A. (1983). Interpenetrating Subsamples. In *Encyclopedia of Statistical Sciences*, 4, S. Kotz and N.L. Johnson (eds), 197–201. New York: Wiley.
- Bailar, B.A., Bailey, L., and Stevens, J. (1977). Measures of Interviewer Bias and Variance. *Journal of Marketing Research*, 14, 337–343.
- Bailey, L., Moore, T.F., and Bailar, B.A. (1978). An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample. *Journal of the American Statistical Association*, 73, 16–23.

- Behrens, K., and Löffler, U. (1999). Aufbau des ADM-Stichproben-Systems. In Stichprobenverfahren in der Umfrageforschung, (ed.), ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. and AG.MA Arbeitsgemeinschaft Media-Analyse e.V., 69–91. Opladen: Leske + Budrich. [In German]
- Belak, E. and Vehovar, V. (1995). Interviewers' Effects in Telephone Surveys. The Case of International Victim Survey. In Contributions to Methodology and Statistics. Methodoloskizvezki 10, A. Ferligoj and A. Kramberger (eds), 85–98. Ljubljana: FDV.
- Biemer, P.P. and Stokes, S.L. (1985). Optimal Design of Interviewer Variance Experiments in Complex Surveys. *Journal of the American Statistical Association*, 80, 158–166.
- Biemer, P.P. and Trewin, D. (1997). A Review of Measurement Error Effects on the Analysis of Survey Data. In *Survey Measurement and Process Quality*, L.E. Lyberg, P.P. Biemer, M. Collins, E. DeLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), 603–632. New York: John Wiley and Sons.
- Cannell C.F. (1953). A Study of the Effects of Interviewers' Expectations Upon Interviewing Results, Ohio State University: Dissertation.
- Collins, M. (1980). Interviewer Variability: A Review of a Problem. *Journal of the Market Research Society*, 22, 77–95.
- Collins, M. and Butcher, B. (1982). Interviewer and Clustering Effects in an Attitude Survey. *Journal of the Market Research Society*, 25, 39–58.
- Cuzick, J. (1985). A Wilcoxon-type Test for Trend. *Statistics in Medicine*, 4, 87–90.
- Davis, P. and Scott, A. (1995). The Effect of Interviewer Variance on Domain Comparisons. *Survey Methodology*, 21, 99–106.
- Fellegi, I.P. (1964). Response Variance and Its Estimation. *Journal of the American Statistical Association*, 59, 1016–1041.
- Fowler, F.J. (1991). Reducing Interviewer-related Error through Interviewer Training, Supervision, and Other Means. In *Measurement Errors in Surveys* P.P. Biemer, R.M. Groves, L.E. Lyberg, N. Mathiowetz, and S. Sudman (eds), 259–278. New York: John Wiley and Sons.
- Fowler, F.J. and Mangione, T.W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park: Sage.
- Gales, K. and Kendall, M.G. (1957). An Inquiry Concerning Interviewer Variability (with Discussion). *Journal of the Royal Statistical Society, Series A*, 120, 121–147.
- Goldstein, H., Browne, W., and Rasbash, J. (2000). Extensions on the Intra-unit Correlation Coefficient to Complex Generalised Linear Multilevel Models. Manuscript, Institute of Education, London, UK.
- Gray, P.G. (1956). Examples of Interviewer Variability Taken from Two Sample Surveys. *Applied Statistics*, 5, 73–85.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley and Sons.
- Groves, R.M. and Magilavy, L.J. (1986). Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. *Public Opinion Quarterly*, 50, 251–266.
- Hagenaars, J.A. and Heinen, T.G. (1982). Effects of Role-independent Interviewer Characteristics on Responses. In *Response Behaviour in the Survey-Interview*, W. Dijkstra and J. van der Zouwen (eds), 91–130. London: Academic Press.

- Hansen, M.H., Hurwitz, W.N. and Bershada, M.A. (1961). Measurement Errors in Censuses and Surveys. *Bulletin of the International Statistical Institute*, 38, 359–374.
- Hanson, R.H. and Marks, E.S. (1958). Influence of the Interviewer on the Accuracy of Survey Results. *Journal of the American Statistical Association*, 53, 635–655.
- Hedges, B. (1980). Discussion of the Paper by Drs Verma and Scott and Mr O’Muircheartaigh. *Journal of the Royal Statistical Society, Series A*, 143, 465–466.
- Hermann, D. (1983). Die Priorität von Einstellungen und Verzerrungen im Interview. Eine Methodenuntersuchung anhand der Daten der Allgemeinen Bevölkerungsumfrage 1980. *Zeitschrift für Soziologie*, 12, 242–252. [In German]
- Hoag, W.J. and Allerbeck, K.R. (1981). Interviewer- und Situationseffekte in Umfragen: Eine log-lineare Analyse, *Zeitschrift für Soziologie*, 10, 413–426. [In German]
- Hyman, H.H., Cobb, W.J., Feldman, J.J., Hart, C.W., and Stember, C.H. (1954). *Interviewing in Social Research*. Chicago: University of Chicago Press.
- Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. *Journal of the American Statistical Association*, 57, 92–115.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- Kish, L. (1995). Methods for Design Effects. *Journal of Official Statistics*, 11, 55–77.
- Kish, L. and Frankel, M.R. (1974). Inference from Complex Samples. *Journal of the Royal Statistical Society, Series B*, 36, 1–37.
- Kreuter, F. (2002). *Kriminalitätsfurcht: Messung und methodische Probleme*. Opladen: Leske + Budrich. [In German]
- Lee, E.S., Forthofer, R.N., and Lorimor, R.J. (1989). *Analyzing Complex Survey Data*. Newbury Park: Sage.
- Mangione, T.W., Fowler, F.J., and Louis, T.A. (1992). Question Characteristics and Interviewer Effects. *Journal of Official Statistics*, 8, 293–307.
- Martin, E. (1983). Surveys as Social Indicators: Problems in Monitoring Trends. In *Handbook of Survey Research*, P.H. Rossi, J.D. Wright, and A.B. Anderson (eds), 677–743. Orlando: Academic Press.
- McPherson, M., Smith-Lovin, L., and Cook, J.M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27, 415–444.
- O’Muircheartaigh, C. (1976). Response Errors in an Attitudinal Sample Survey. *Quality and Quantity*, 10, 97–115.
- O’Muircheartaigh, C. and Campanelli, P. (1998). The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision. *Journal of the Royal Statistical Society, Series A*, 161, 63–77.
- Pickery, J. and Loosveldt, G. (2001). An Exploration of Question Characteristics that Mediate Interviewer Effects on Item Nonresponse. *Journal of Official Statistics*, 17, 337–350.
- Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable Estimation of Generalized Linear Mixed Models Using Adaptive Quadrature. *The Stata Journal*, 2, 1–21.
- Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Longford, I., and Lewis, T. (2000). *A User’s Guide to MLwiN*. University of London: Institute of Education.

- Rodríguez, G. and Goldman, N. (1995). An Assessment of Estimation Procedures for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society, Series A*, 158, 73–89.
- Rice, S.A. (1929). Contagious Bias in the Interview: A Methodological Note. *American Journal of Sociology*, 35, 420–423.
- Rustemeyer, A. (1977). Measuring Interviewer Performance in Mock Interviews. *Proceedings of the American Statistical Association, Social Statistics Section*, 341–346.
- Rässler, S. and Schnell, R. (2004). Multiple Imputation for Unit-Nonresponse versus Weighting including a Comparison with a Nonresponse Follow-Up Study. Discussion paper 65/2004, Nürnberg. <http://www.uni-konstanz.de/FuF/Verwiss/Schnell/Basel2004.pdf>
- Scheuch, E.K. (1974). Auswahlverfahren in der Sozialforschung. In *Handbuch der empirischen Sozialforschung Bd. 3a*, R. König (ed.), 1–96. Stuttgart: Enke. [In German]
- Schnell, R. (1997). Nonresponse in Bevölkerungsumfragen. Opladen: Leske + Budrich, [In German]
- Schnell, R. and Kreuter, F. (2000). Das DEFECT-Projekt: Sampling-Errors und Nonsampling-Errors in komplexen Bevölkerungsstichproben, *ZUMA-Nachrichten*, 47, 89–101. [In German] [www.gesis.org/Publikationen/Zeitschriften/ZUMA\\_Nachrichten/documents/pdfs/zn47\\_10-mitteilungen.pdf](http://www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten/documents/pdfs/zn47_10-mitteilungen.pdf).
- Schnell, R. and Kreuter, F. (2001). Neue Software-Werkzeuge zur Dokumentation der Fragebogenentwicklung, *ZA-Information*, 48, 56–70. [In German] [www.za.uni-koeln.de/publications/pdf/za\\_info/ZA-Info-48.pdf](http://www.za.uni-koeln.de/publications/pdf/za_info/ZA-Info-48.pdf).
- Shao, J. (1996). Resampling Methods in Sample Surveys. Invited paper, *Statistics*, 27, 203–237, with discussion, 237–254.
- Van Tilburg, T. (1998). Interviewer Effects in the Measurement of Personal Network Size. *Sociological Methods and Research*, 26, 300–328.
- Tucker, C. (1983). Interviewer Effects in Telephone Surveys. *Public Opinion Quarterly*, 47, 84–95.

Received July 2003

Revised October 2004