

# MODELING GROUP-SPECIFIC INTERVIEWER EFFECTS ON SURVEY PARTICIPATION USING SEPARATE CODING FOR RANDOM SLOPES IN MULTILEVEL MODELS

JESSICA M. E. HERZING 

ANNELIES G. BLOM 

BART MEULEMAN

Despite its importance in terms of survey participation, the literature is sparse on how face-to-face interviewers differentially affect specific groups of sample units. This paper demonstrates how an alternative parametrization of the random components in multilevel models, so-called separate coding, delivers valuable insights into differential interviewer effects for specific groups of sample members. In the example of a face-to-face recruitment interview for a probability-based online panel, we detect small interviewer effects regarding survey participation for non-Internet households, whereas we find sizable interviewer effects for Internet households. We derive practical guidance for survey practitioners to address differential interviewer effects based on the proposed variance decomposition.

**KEYWORDS:** Hierarchical model; Interviewer variance; Mixed-effects model; Nonresponse; Online panel.

JESSICA M. E. HERZING is a Scientific Collaborator at the ICER, Faculty of Social and Political Science, Department of Social Science, University of Bern, ICER, Fabrikstrasse 8, 3012 Bern, Switzerland and Data Science, School of Social Sciences, University of Mannheim, Mannheim, Germany. ANNELIES G. BLOM is a Professorship in Data Science at the Data Science, School of Social Sciences, University of Mannheim, 68131 Mannheim, Germany and DIGSSCORE, Department of Administration and Organization Theory, University of Bergen, Bern, Switzerland. BART MEULEMAN is Professor at the Faculty of Social Science, Centre for Sociological Research, University of Leuven, Parkstraat 45 - box 3601, 3000 Leuven, Belgium

The authors disclosed receipt of the following financial support for the research, authorship, and publication of this article: The GIP is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through the Collaborative Research Center (SFB) 884 “Political Economy of Reforms” (SFB 884; Project-ID 139943784).

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

## Statement of Significance

This paper contributes to the literature on detecting interviewer effects and explaining interviewer effects. Applying an alternative parametrization of the random components in multilevel models, so-called separate coding, we deliver novel insights into differential interviewer effects for specific groups of sample members. By differentiating interviewer effects across different sample groups, we also speak to the literature on survey representativeness and interviewer-induced sample imbalances. Finally, we provide survey practitioners with practical guidance to address differential interviewer effects and a hands-on approach to identifying underperforming interviewers.

## 1. INTRODUCTION

In face-to-face surveys, interviewers play a crucial role in recruiting the respondent sample. Differences in interviewers' success at recruiting respondents are observed in the form of interviewer effects (Couper and Groves 1992; Hox and de Leeuw 2002; Durrant and Steele 2009; Durrant et al. 2010; West and

\*Address correspondence to Jessica M. E. Herzing, Department of Social Science, University of Bern, ICER, Fabrikstrasse 8, 3012 Bern, Switzerland; E-mail: jessica.herzing@unibe.ch.

Most GIP survey data are available to the scientific community as Scientific Use Files (SUFs) from the GIP data archive.<sup>1</sup> In addition, data that cannot be published in the archive according to the EU General Data Protection Regulations (EU-GDPR) analyses can be accessed at the On-Site Data Access (ODA) facilities of the GIP Secure Data Center (SDC) located at the Collaborative Research Center "Political Economy of Reforms" (SFB 884), University of Mannheim, B6 30-32 Mannheim, Germany. Researchers wishing to use the ODA, please get in touch with GIPinfo@uni-mannheim.de. The Collaborative Research Center is funded until September 2022. After that, the ODA will move to a new SDC. If you wish to access the ODA after September 2022 or wish to access the software code, please get in touch with the corresponding author for further information.

1. <https://dbk.gesis.org/dbksearch/GDesc2.asp?no=0109&tab=&ll=10&notabs=1&db=E> accessed August 2021.

This paper stems from Jessica M. E. Herzing's doctoral dissertation, supervised by Annelies Blom and Frauke Kreuter, she thanks them for their helpful comments in developing this research. Jessica also thanks the additional members of her dissertation committee Florian Keusch and Edith de Leeuw, for additional comments. We are grateful to Daniela Ackermann-Piek and Oliver Lipps for their comments on previous drafts of the paper. Furthermore, we would like to thank Klaus Pforr, Michael Ochsner, Kristen Olson, Jolene Smyth, Brady West, the participants of the Interviewer Workshop 2018, the members of the MLM email list, and the *JSSAM* editor, Kristen Olson, an anonymous associate editor, and three anonymous reviewers for valuable suggestions and lively discussions. The authors gratefully acknowledge funding for their research by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through the Collaborative Research Center (SFB) 884 "Political Economy of Reforms" (SFB 884; Project-ID 139943784), in particular of project Z1 (German Internet Panel, GIP). Since this study uses the GIP data as a secondary data source, its study design and analysis were not preregistered.

Blom 2017). Moreover, the interviewers' behavior and characteristics determine the extent to which population subgroups respond to the survey request and, thus, their representation in the respondent sample (West and Olson 2010; West et al. 2013; Loosveldt and Beullens 2014; West 2020).

For some surveys, researchers can point to specific population subgroups that they are particularly afraid to underrepresent. For example, education surveys are in danger of underrepresenting functionally illiterate subgroups because illiterate persons may feel ashamed of demonstrating their illiteracy to the interviewer (for instance, Helmschrott and Martin 2014). Similarly, online surveys are in danger of underrepresenting persons who do not have access to or the skills to operate computers and/or the Internet, even if the researchers provide them with the necessary devices and support (Eckman 2016; Revilla et al. 2016; Cornesse and Schaurer 2021). In face-to-face surveys, interviewers can be more or less successful at achieving interviews with such hard-to-reach subgroups.

Research into interviewer effects on survey participation commonly uses multilevel models (also known as hierarchical or mixed-effects models) to estimate the amount of between-interviewer variance (level 1: sample units, level 2: interviewers). In its standard parametrization (so-called contrast coding), the variance components of a multilevel model with random slopes deliver insights into the between-interviewer variance, that is, into overall interviewer effects on survey participation of specific population subgroups. However, this parameterization does not address the question of how large the interviewer variances are separately for different population subgroups.

Attempts to address this question have been made in previous studies. Loosveldt and Beullens (2014), for example, incorporated respondent characteristics by adding dummy variables in the random part of the model (contrast coding). Unfortunately, this parameterization is not satisfactory because the interpretation of the variance components is not intuitive and potentially even misleading: The obtained random slope variance for the dummies only indicates to what extent the *difference* in response rates between sample groups varies across interviewers (see also West and Elliott 2014). Beullens et al. (2019) and Loosveldt and Wuyts (2020) also aim to identify differential interviewer effects in surveys, albeit in terms of interviewer measurement variances during the interview. In a two-step procedure, they extend the basic multilevel model with a conditional random interviewer effect model to estimate the effect of respondent characteristics on the variability of intraclass correlations.

Conversely, our method identifies to what extent interviewers differ in their success in recruiting a specific subgroup as opposed to another subgroup. In the alternative parameterization (so-called separate coding) of the multilevel model (Jones 2013) that we propose in this paper, the variance components reveal whether the size of the interviewer effect on the participation of one subgroup differs significantly from the size of the interviewer

effect on another subgroup. Furthermore, it allows us to observe interviewer characteristics associated with survey participation in one subgroup separately from interviewer characteristics associated with participation in another subgroup. In previous research, the proposed parameterization of the multilevel interviewer effects model has been used to investigate differences in interviewer effects between groups of interviewers who used different interviewing techniques (West and Elliott 2014; West et al. 2018a, 2018b, 2022). This paper adds to this literature by using this parametrization approach to explain differential interviewer variances within respondent subgroups using interviewer characteristics. Our parametrization of the multilevel interviewer effects model enables us to investigate the following research questions: Does the size of interviewer effects on survey participation vary across specific sample groups? And which interviewer characteristics explain survey participation in one sample group and which in other sample groups?

To illustrate the two parameterizations in the context of interviewer effects, we use data from the face-to-face recruitment interview of the probability-based German Internet Panel (GIP), a data collection characterized by group-specific survey participation. Response rates among online (persons with computer and/or Internet access) are significantly higher than response rates among offline (persons without computer and/or Internet access), even though all offline were offered equipment and support to enable their participation in the online panel (Blom et al. 2017; Herzing and Blom 2019). The question then arises whether this is solely an effect of the sample units themselves or whether some interviewers are particularly good or bad at convincing offline or online to participate in the panel. Hence, we ask: Can we identify interviewer effects specifically for the underrepresented group of offline, and if so, can we explain these with known interviewer characteristics? With this information, the GIP may subsequently address this imbalance in survey participation at the stage of the face-to-face recruitment interviews. For example, they may send out those interviewers to the offline that have the specific interviewer characteristics associated with higher participation rates amongst offline.

In summary, this paper addresses gaps in the literature related to explaining differential interviewer variances within respondent subgroups using interviewer characteristics by tackling four research questions:

- (1) How can multilevel models be parameterized such that they reveal and explain differences in interviewer effects between sample groups?
- (2a) To what extent do interviewers affect survey participation among online and offline in the GIP?
- (2b) Does the size of the interviewer effect on survey participation differ between online and offline in the GIP?

- (2c) If so, which interviewer characteristics are associated with interviewer effects on survey participation differentially across the two sample groups?

## 2. PREVIOUS RESEARCH INTO EXPLAINING INTERVIEWER EFFECTS ON SURVEY PARTICIPATION IN FACE-TO-FACE SURVEYS

Most research into interviewer effects on survey participation relies on the standard parametrization of the multilevel model. With such multilevel models, a multitude of studies has detected interviewer variances in survey participation rates (see [West and Blom 2017](#)). These multilevel models take the clustering of sampling units into interviewers into account by adding a variance component at the interviewer level. Overall, interviewers tend to increase response in surveys compared to surveys where no interviewers are present, for example, web surveys (see [Heerwegh 2009](#)). However, not all interviewers are equally successful at getting sample units to participate ([Couper and Groves 1992](#); [Hox and de Leeuw 2002](#); [Durrant and Steele 2009](#); [Durrant et al. 2010](#); [West and Blom 2017](#)). If interviewers systematically differ in their success in obtaining responses from different sample groups, they may introduce nonresponse error variance or bias ([West and Olson 2010](#); [West et al. 2013](#); [Loosveldt and Beullens 2014](#)).

Several academic studies identified interviewer-level predictors that explain the interviewer variances in survey participation (e.g., [Blom et al. 2011](#); [Jäckle et al. 2012](#); [Ackermann-Piek et al. 2020](#)). While the majority found that socio-demographic characteristics of interviewers have no or only small effects on the propensity to respond, interviewer experience, and positive interviewer behaviors and attitudes are predictive of an interviewer's success ([West and Blom 2017](#)).

Combining this academic interest with an objective to provide specific practical recommendations, researchers have proposed adaptations of the standard multilevel model to identify specific groups of survey interviewers that produce particularly high interviewer-related variance (e.g., [Lipps and Pollien 2011](#); [Brunton-Smith et al. 2012](#); [West and Elliott 2014](#); [West et al. 2018a](#)). Identifying these specific interviewer groups may enable survey operators to approach them specifically for retraining.

Another important approach to interviewer effects on survey participation in face-to-face surveys is the decomposition of the total interviewer effects observed in survey data into the effects interviewers have during the recruitment process (i.e., on response rates) and the effects they have on the interviewing process (i.e., on measurement). Such studies necessitate validation data, which unfortunately are seldom available. This high demand for the data source and the complexity of the statistical modeling limits the number of investigations

into the issue, with two notable exceptions: [West et al. \(2018a\)](#) and [West and Olson \(2010\)](#).

Both of these adaptations of the multilevel model deliver valuable pieces of the puzzle of interviewers' impact on survey data collection. Our approach adds another important piece. In contrast to the decomposition of measurement and survey participation effects, our approach focuses on survey participation only. Consequently, it is less data hungry and therefore provides scope for more widespread practical implementation. In contrast to the focusing on groups of *interviewers* that share characteristics and are likely to underperform (e.g., inexperienced interviewers), our approach focuses on groups of *sample units* that share characteristics and are likely to be difficult to recruit into a survey (e.g., functionally illiterate or non-Internet sample units). As a consequence, it provides scope for countering expected sample imbalances by deploying the most suitable interviewers to such groups of sample units.

### 3. PARAMETERIZATION OF INTERVIEWER EFFECTS FOR SPECIFIC SAMPLE GROUPS

In the following, we describe the two parameterization strategies for categorical grouping variables in two-level interviewer effect models with random slopes. In any multilevel model, we can code categorical predictors either by contrast coding or separate coding (see [Verbeke and Molenberghs 2000](#), chap. 12.1). In contrast coding, we include all but one category of the predictor variable in the model plus an intercept. The omitted category forms the reference group in relation to which all other groups are interpreted. This parametrization delivers insights into between-interviewer variances and thus informs about the presence and overall magnitude of interviewer effects. However, this parameterization does *not* address the presence and magnitude of the interviewer effects *separately* for each predictor category. Therefore, in separate coding ([Jones, 2013](#)), we include all categories of the predictor variable in the model, and the intercept is omitted. This parametrization allows a new interpretation of the variance components and, consequently, answers new research questions (see, e.g., [West and Elliott 2014](#); [West et al. 2018b, 2022](#)).

#### 3.1 Parameterization with Contrast Coding

In contrast coding (also referred to as reference group or dummy coding), the average value for the reference group is captured by the intercept, while the scores of other categories are estimated as a deviation from this reference.

To illustrate this parametrization, we consider a two-level logistic regression model on survey participation with sample units on level 1 and interviewers on

level 2 (i.e., sample units nested in interviewers). The survey participation of sample unit  $i$ , who is interviewed by interviewer  $j$ , is denoted by  $\text{parti}_{ij}$  as

$$\text{parti}_{ij} = \begin{cases} 0 & \text{no survey participation,} \\ 1 & \text{survey participation.} \end{cases}$$

As in single-level logistic regressions, the probability  $\pi$  of observing the value 1 in the dichotomous variable  $\text{parti}_{ij}$  is modeled as a logistic transformation resulting in

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right)$$

with  $\pi_{ij}$  denoting the response probability of sample unit  $i$ , who is approached by interviewer  $j$ . Furthermore, we consider a variable that distinguishes groups of sample units with different probabilities to respond (the variable has to be available on the sampling frame or from a prior wave of data collection), for example, high- versus low-educated persons or offline versus online. Two sample groups are introduced as a dummy predictor ( $\delta^1$ ) coded

$$\delta^1 = \begin{cases} 0 & \text{sample group A} \\ 1 & \text{sample group B.} \end{cases}$$

Next, we take into account that the response probability of sample groups A and B can vary across interviewers, and we add a random slope for the dummy variable  $\delta^1$ . The resulting multilevel model is formalized in (1)–(3).

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_{0j} + \beta_{1j}\delta_{ij}^1 \quad (1)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (3)$$

By substituting (2) and (3) into (1), we obtain the model in reduced form:

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \gamma_{00} + \gamma_{10}\delta_{ij}^1 + u_{1j}\delta_{ij}^1 + u_{0j} \quad (4)$$

with

$$u_{0j} \sim N(0, \sigma_{u0}^2); \quad u_{1j} \sim N(0, \sigma_{u1}^2) .$$

This parameterization with contrast coding has important consequences for the interpretation of the parameters. In this model, parameters  $\gamma_{00}$  and  $\gamma_{10}$  are

the fixed effects.  $\gamma_{00}$  is the grand intercept, representing the logit transformation of the probability of participation for group A (i.e., the reference category) across all interviewers.  $\gamma_{10}$  captures how the logit to respond differs for sample group B compared to sample group A, again on average across all interviewers. The variation between interviewers is incorporated in the random part of the model. Random intercept  $u_{0j}$  denotes how the response rate a particular interviewer  $j$  obtains among sample group A (the reference category) deviates from that of the average response among this group. Therefore, the random intercept variance  $\sigma_{u0}^2$  represents the cross-interviewer variation in the recruitment success, specifically for respondents of group A. The random slope  $u_{1j}$  represents the difference in obtained participation between sample groups A and B for a particular interviewer  $j$ . The random slope variance  $\sigma_{u1}^2$ , therefore, captures how the *difference* in participation between respondents of both groups varies across interviewers (as stated in our research question 2a). Thus, with contrast coding, the interpretation of the random slope variance is thus not very intuitive and does *not* directly provide insight into the size of interviewer effects *for each sample group*.

### 3.2 Parameterization with Separate Coding

Research questions 1, 2b, and 2c address the *size* of the interviewer effect for group A, the size of the interviewer effect for group B, whether these interviewer effects differ significantly from each other, and the predictors of these separate interviewer effects. To answer these research questions, we need a parametrization of the multilevel model with separate coding.

In this parametrization, all dummies of a categorical variable are included in the model without omission. However, no intercept is included because this would lead to perfect multicollinearity. As a result, each binary variable represents a direct estimate of the group mean.

We adapt the parameterization of model (4). We switch from contrast coding in both the fixed and the random parts of the multilevel model to a model that retains contrast coding for the fixed part but uses separate coding in the random part. This means that in the random part, two dummies with random slopes are introduced ( $u_{1j}$  for group A and  $u_{2j}$  for group B), and the random intercept ( $u_{0j}$ ) is omitted (see Jones 2013, pp. 136–8). The model with contrast coding in the fixed part and separate coding in the random part is formalized as

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \gamma_{00} + \gamma_{10}\delta_{ij}^1 + u_{1j}\delta_{ij}^1 + u_{2j}\delta_{ij}^2 \quad (5)$$

with

$$u_{1j} \sim N(0, \sigma_{u1}^2); \quad u_{2j} \sim N(0, \sigma_{u2}^2)$$



where

$$\delta^2 = \begin{cases} 0 & \text{sample group B,} \\ 1 & \text{sample group A.} \end{cases}$$

Note that the fixed part of (5) is identical to the fixed part of (4). The random slope variances of the two dummies reveal the size of the interviewer effect for each group separately. Thus, the random slope variance  $\sigma_{u1}^2$  captures how the difference in response rates between sample units of group A varies across interviewers, whereas the random slope variance  $\sigma_{u2}^2$  captures how the difference in response rates between sample units of group B varies across interviewers.

To gain more insight into group-specific interviewer effect sizes, we can estimate whether the size of the interviewer effect for group A differs significantly from the size of the interviewer effect for group B using a Wald test. However, variances are defined as non-negative and, hence, the null hypothesis is on the boundary of the parameter space, i.e., the standard errors and confidence intervals are not meaningful. Nevertheless, the significance of the random slopes  $u_{1j}\delta_{ij}^1$  and  $u_{2j}\delta_{ij}^2$  can be assessed following an approach by Molenberghs and Verbeke (2007), who suggest testing the deviance difference between the models ( $D_0 - D_1$ ) against a 50–50 mixture  $\chi_p^2$  and  $\chi_{p+1}^2$  distribution for two-sided hypothesis tests in unconstrained multilevel models (see also Snijders and Bosker 2011, pp. 98–9), when the deviance difference is larger than 0. In case the deviance difference is equal to 0, the random slope variance is not significant.

A further possibility consists of testing whether two variance components are significantly different (e.g., to investigate whether the size of interviewer effects is different between two groups of respondents). Such a test requires a different statistical approach and is discussed in West (2020, p. 330). Although both parameterizations—contrast and separate coding—yield different insights, it is important to stress that both models are statistically equivalent. This means that we can transform the covariance and variances from (4) into the variance components of (5) (for further details see Rabe-Hesketh and Skrondal 2008, chap. 11.4). This statistical equivalence between models (4) and (5), however, only holds if the covariance matrix for the random effects in model (5) is specified as unstructured, thus, allowing a correlation between both random effects at the interviewer level (see Rabe-Hesketh and Skrondal 2008, chap. 11.4). The resulting variance–covariance matrix for the random slopes  $u_{1j}$ ,  $u_{2j}$  is given by

$$\delta \begin{bmatrix} u_{1j} \\ u_{2j} \end{bmatrix} = \begin{bmatrix} \sigma_{u1}^2 & \sigma_{u1u2} \\ \sigma_{u1u2} & \sigma_{u2}^2 \end{bmatrix}$$

### 3.3 Parameterization with Cross-Level Interactions

To assess whether particular interviewer characteristics affect survey participation differently per sample group (research question 2c), we augment the random slope model with separate coding by including cross-level interactions between the sample group indicator and the interviewer characteristics. The model in (5) is thus extended by adding interviewer characteristic  $Z_j$  as well as an interaction between  $Z_j$  and  $\delta_{ij}^1$  (i.e., the dummy for sample group):

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \gamma_{00} + \gamma_{10}\delta_{ij}^1 + \gamma_{11}\delta_{ij}^1 Z_j + \gamma_{01}Z_j + u_{1j}\delta_{ij}^1 + u_{2j}\delta_{ij}^2 \quad (6)$$

with

$$u_{1j} \sim N(0, \sigma_{u1}^2); \quad u_{2j} \sim N(0, \sigma_{u2}^2).$$

Note that, since the cross-level interaction is located in the fixed part of the model, we use contrast coding to model the interplay between the sample group and the interviewer characteristic.  $\gamma_{01}$  is the parameter for the main effect for  $Z_j$  and thus represents the impact of  $Z_j$  on the logit to respond, conditional on  $\delta_{ij}^1$  being equal to 0. In other words,  $\gamma_{01}$  captures how the interviewer variable affects the propensity to respond for group A (the reference category). Furthermore, interaction parameter  $\gamma_{11}$  indicates how the impact of  $Z_j$  in sample group B deviates from the effect in the reference category. This parameterization thus allows us to ascertain whether the interviewer-assisted mechanisms driving (non)response differ between sample groups.

## 4. PRACTICAL GUIDANCE FOR SURVEY PRACTITIONERS

This section guides survey practitioners in the use of multilevel models with separate coding in the random part. We provide interpretations of various scenarios of estimates that may occur when modeling interviewer effects in this way (see table 1).<sup>2</sup>

Scenario 1: Both interviewer effects are larger than 0 ( $\hat{\sigma}_{u1}^2 > 0$  and  $\hat{\sigma}_{u2}^2 > 0$ ) and we find a difference in the size of the interviewer effects ( $\hat{\sigma}_{u1}^2 > \hat{\sigma}_{u2}^2$ ). In such a situation, interviewers might well be the reason for differences in the response rates with respect to the sample group examined. To tackle the differences in the interviewer variances, survey practitioners

2. Note that adding random slopes to any multilevel model necessitates a significant test of test of deviance differences.

Table 1. Three Scenarios of Potential Outcomes and Their Practical Implications

Scenario 1	Scenario 2	Scenario 3
Interviewer effect size		
$\hat{\sigma}_{u1}^2 > 0$	$\hat{\sigma}_{u1}^2 > 0$	$\hat{\sigma}_{u1}^2 > 0$
$\hat{\sigma}_{u2}^2 > 0$	$\hat{\sigma}_{u2}^2 > 0$	$\hat{\sigma}_{u2}^2 \triangleq 0$
Differences in interviewer effect sizes		
$\hat{\sigma}_{u1}^2 > \hat{\sigma}_{u2}^2$	$\hat{\sigma}_{u1}^2 \triangleq \hat{\sigma}_{u2}^2$	$\hat{\sigma}_{u1}^2 > \hat{\sigma}_{u2}^2$
Outcome		
Interviewer effects are present for groups A and B. There are significant differences in interviewer effect sizes between groups A and B.	Interviewer effects are present for both groups. However, there are no significant differences in interviewer effect sizes between groups A and B.	Interviewer effects are present for group A but not for group B. Significant differences in interviewer effect sizes between groups A and B.
Interpretation		
Interviewers may be the cause of the differences in response rates between sample groups.	It is unlikely that interviewers are the cause of the differences in response rates between sample groups.	Interviewers may be the cause of the differences in response rates between sample groups.
Next steps		
Investigate whether (un)successful <sup>a</sup> interviewers are equally distributed.	Investigate whether there are other explanations for differences in response rates, like sample composition effects.	Investigate whether (un)successful interviewers have the same characteristics across sample groups and whether (un)successful interviewers are equally distributed.

NOTE.— <sup>a</sup>Successful or unsuccessful interviewers are identified by investigating interviewers’ individual random slopes or response rates per sample group.

need to investigate whether (un)successful interviewers are equally distributed across the sample groups. In the case of unequal distributions, the simplest remedy is to equally distribute the more and the less successful interviewers across the sample groups to reduce imbalances. If interviewers are unequally successful at recruiting one group compared to another, including specific training elements addressing recruitment strategies for the more difficult sample group in the (re-)training of the interviewers struggling with this group should reduce and/or equalize the interviewer variances across groups (see Groves and McGonagle 2001). In particular, the training should aim to reduce the interviewer effect in the sample group with more substantial interviewer effects (in our example, group A;  $\hat{\sigma}_{u1}^2$ ).

Scenario 2: We find interviewer effects for both sample groups ( $\hat{\sigma}_{u1}^2 > 0$  and  $\hat{\sigma}_{u2}^2 > 0$ ); however, there is no significant difference in the sizes of these interviewer effects ( $\hat{\sigma}_{u1}^2 \triangleq \hat{\sigma}_{u2}^2$ ). In such a situation, it is unlikely that interviewers are the reason for imbalances in response rates. Therefore, there are likely other explanations for the differences in response rates across the two sample groups. For example, the sample composition within interviewers may differ, leading to some interviewers handling sampling units with lower response propensities. This phenomenon is frequently observed in practice (see, e.g., Blom 2012).

Scenario 3: We find interviewer effects for one sample group ( $\hat{\sigma}_{u1}^2 > 0$ ) but not for the other ( $\hat{\sigma}_{u2}^2 \triangleq 0$ ). Furthermore, the difference in the interviewer effects is significant ( $\hat{\sigma}_{u1}^2 > \hat{\sigma}_{u2}^2$ ). In such a situation, interviewers are likely the reason for response rate imbalances across sample groups. To tackle the differences in the interviewer variances, survey practitioners need to investigate whether (un)successful interviewers have the same characteristics across sample groups, that is, whether it is the same kind of interviewers that approach the two groups of sample units. In addition, survey practitioners need to investigate whether (un)successful interviewers are equally distributed across the sample groups. In the case of unequal distributions, the more and the less successful interviewers should be redistributed across the sample groups to reduce imbalances and/or interviewers that are less successful with one sample group should be approached for (re)training.

Our proposed modeling allows identifying exactly which interviewers are (un)successful at recruiting certain sample groups. How to do this graphically, for example, is discussed further below.

Finally, the covariance ( $\hat{\sigma}_{u1,u2}$ ) may tell us something about the interviewer effects in case of a separate coding of the random slopes. A significant positive covariance between the random slope for sample group A and the random slope for sample group B, for example, means that interviewers who are good at gaining response from sample group A are also good at gaining response from sample group B. Conversely, a significant negative covariance means that interviewers who are good at gaining response from one group are bad at gaining response from the other. Again, such findings may be leveraged to effectively support interviewers' fieldwork.

On a final note, using specific types of interviewers for specific types of cases is costly and often restricted due to travel constraints, project budgets, etc., whereas careful (re)training is likely to be cheaper and easier to implement.

## 5. APPLICATION

The following application showcases how to separate coding for the random part of an interviewer effects model can deliver insights into differential

interviewer effects for specific sample groups. For this purpose, we examine data from the face-to-face recruitment interview of the probability-based online panel GIP.

### 5.1 The GIP Data

Initially set up in 2012, the GIP was the first probability-based online panel of the general population in Germany and one of the first probability-based online panels worldwide that included the offline population (see [Blom et al. 2015](#)). Well aware of the challenges faced when aiming to cover the general population, the GIP endeavored to limit selectivities, in particular, biases due to non-response and noncoverage in the online mode.

To optimize representativeness, the GIP took various measures, of which the most important ones are listed in the following. First, the GIP sample is based on a strict random probability sample of the general population aged 16–75 years at recruitment. The sample was drawn as a three-stage area probability sample. At the first sampling stage, a random sample of areas was drawn from a database that covers all areas in Germany. Within each primary sampling unit (PSU), lists recorded every household along a predefined random route until they had listed 200 households. At the second stage, a random sample of households was drawn for a 15-minute face-to-face interview. The face-to-face interview identified the age-eligible household members, who, at the third stage, were all invited to the online panel ([Blom et al. 2015](#)). Persons living in households without a broadband Internet connection and/or computer (so-called offliners) were informed that they were also invited to participate in the online study and that they would receive equipment to enable their participation (see [Herzing and Blom 2019](#)).

Our application uses data from the initial 2012 GIP recruitment and from the 2014 GIP refresher sample (face-to-face response rate 2012: 52.1 percent [AAPOR RR2]; 2014: 47.5 percent [AAPOR RR1]; details on population, response rates, and fieldwork can be found in [Blom et al. 2015, 2017](#) and on the GIP website<sup>3</sup>). In a study on the added value of offline households in the GIP sample, [Blom et al. \(2017\)](#) found that “[...] despite the careful survey design, response rates were significantly higher among previously online than among previously offline sample units” (p. 503). We thus investigate whether these differences in response rates between onliners (in the following equivalent to sample group A in the equations) and offliners (in the following equivalent to sample group B in the equations) are due to the interviewers and, if so, in which way. Previous research suggests having access to computers and/or the Internet strongly correlates with key sociodemographic variables ([Blom et al. 2017; Herzing and Blom 2019](#)). As a consequence, the underrepresentation of

3. <https://www.uni-mannheim.de/en/gip/for-data-users/questionnaires-and-documentation/> accessed August 2021.

offliners puts social research based on the GIP data at risk for biased estimates (see the common cause model, [Groves 2006](#)).

We base our estimations on all cases of the face-to-face interviews (gross sample) and model interviewer effects on participation in the online panel (operationalized by agreement to participate in the online panel during the interview). In total, 324 interviewers interviewed 5,238 age-eligible respondents during the face-to-face recruitment interviews. Of these respondents, 3,842 respondents agreed to participate in the online panel (2,970 onliners and 872 offliners, for summary statistics of the respondent characteristics, see [appendix A, table A.1 in the supplementary data online](#)).

In the GIP, sample units were not randomly allocated to interviewers in an interpenetrated design (see [West and Blom 2017](#) for details). Furthermore, any region (PSU) was assigned to one interviewer and the vast majority of interviewers were assigned to only one region. Therefore, we cannot statistically disentangle area from interviewer effects. However, to account for differences in the sample composition of the interviewers' assignments, we control for the sample units' age, gender, household size, level of education, employment status, frequency of Internet use, frequency of media consumption, and whether they voted in the last general election (for an English translation of the survey questionnaires see [appendix B in the supplementary data online](#)).<sup>4</sup> In addition, we account for the nonrandom allocation of interviewers by stepwise introducing the interviewer characteristics to the sample unit characteristics for which the areas could have a differential composition (see model 0 and 1 in the [appendix A, tables A.3 and A.4 in the supplementary data online](#) and for further information; [Hox 1994](#); [Blom et al. 2011](#), p. 367; [Steele and Durrant 2011](#); [Loosveldt and Wuyts 2020](#)).

A small proportion of missing values on the sample units' age (< 1 percent) was imputed using predictive mean matching (see, [Little 1988](#); [Morris et al. 2014](#)). The analyses are presented unweighted since we aim to infer interviewer behavior (within sample estimation) rather than the general population. Furthermore, sensitivity analyses showed no effect of the sampling design weights (which included regional clusters) on our estimates.

## 5.2 The Interviewer Data

We augment the GIP data with data from an interviewer survey conducted during the interviewer training. This paper-and-pencil survey covers topics on interviewers' own behavior, interviewers' experience with measurements, interviewers' expectations, interviewers' computer and Internet usage, and interviewers' sociodemographic characteristics (interviewer survey adapted

4. None of the predictors in the model were centered. Therefore, the intercept denotes the propensity to respond for onliners scoring 0 on all control variables.

from Blom and Korbmacher 2013; see appendix A, table A.2 and appendix C in the supplementary data online).

In designing the questionnaire and selecting relevant interviewer characteristics for our models, we followed the literature on factors explaining interviewer effects on nonresponse (see West and Blom 2017 for an overview). Our models tested five interviewer-level indicators measuring interviewers' reported interviewing behavior and work experience (Ackermann-Piek et al. 2020), their habit of tailoring the questionnaire to respondents' needs, their expectations regarding online panel participation rates, and interviewers' ability to explain to offliners how they are supported in their participation in the online panel (Blom and Korbmacher 2013). Unfortunately, we were not able to test for interviewer-respondent liking effects (Durrant et al. 2010) regarding the onliner/offliner status, because all interviewers were onliners per job requirement. (Their job demands them to regularly go online with their laptop to transfer interview data back to the survey agency.)

In total, 274 interviewers completed the interviewer questionnaire in 2012 and 2014. We identified 57 interviewers that participated in both surveys, but for reasons of data protection, we were not allowed to match interviewers across recruitment rounds. Hence, we excluded these interviewers from our analyses. For eleven interviewers, a few missing values were imputed by predictive mean matching (less than 1 percent). Furthermore, we found twenty-eight interviewers who did not interview any offliners and two interviewers that did not interview any onliners in the recruitment interviews. To test the sensitivity of the model, we ran the analysis without these thirty interviewers and found no notable differences in estimations in the variables between the two models (see appendix A, table A.3 and A.4 in the supplementary data online).

## 6. RESULTS

Our application follows the analytical steps set out in section 2. We estimate several two-level logistic regression models (sample units nested within interviewers) with response to the online panel as the dependent variable and with a categorical variable indicating whether a sample unit was an onliner or an offliner as the key independent variable.

To estimate the size of the interviewer effect in our analyses, we commence with a null model, which only controls for respondent characteristics to account for differences in interviewer assignments (see model 0 in appendix A, table A.3 in the supplementary data online). For this model, we identify an

intraclass correlation  $\left( = \frac{\hat{\sigma}_{u0}^2}{\hat{\sigma}_{u0}^2 + \frac{\sigma_{\epsilon}^2}{3}} \right)$  of 25 percent. Thus, a considerable 25 percent of the overall variance in the online panel response probability is located at the interviewer level.

## 6.1 Contrast coding

To illustrate the difference between the two parametrization strategies and to strengthen our argument in favor of a parametrization with separate coding in the random slopes, we first present a model with contrast coding in the random slopes (estimation equivalent to (4), which corresponds to research question 2a). Hence, we estimate a model that includes several control variables for sample units and interviewers, a dummy for offliners (sample unit level), and a random slope for this dummy (model 1 in table 2).

We find that the random intercept variance at the interviewer level ( $\hat{\sigma}_{u0}^2$ ) is significantly different from 0 (based on the test of deviance difference:  $D_0 - D_1 = 2,525.73 - 2,508.48 = 17.22$ ,  $p < 0.001$ ,  $D_0$  presented in appendix A, table A.3 in the supplementary data online). This means that interviewers differ in their success at recruiting onliners (the reference category) into the online panel. The variance ( $\hat{\sigma}_{u1}^2$ ) of the distribution of the interviewer-level slopes of being offline is significant. However, the interpretation of this significant random slope variance is not straightforward: It indicates that there is variation between interviewers with respect to the *difference* in their success in recruiting onliners and offliners (addressing research question 2a). The interpretation of the covariance between random intercepts and slopes ( $\hat{\sigma}_{u0,u1}$ ) is not straightforward to interpret in this context of contrast coding (for further explanations, see Hox 2010, p. 18).<sup>5</sup>

## 6.2 Separate Coding

In the second model in table 2, we use contrast coding in the fixed part and separate coding in the random part (estimation equivalent to (5)). As model 1 and model 2 are statistically equivalent, their interpretation remains the same. However, the parameterization with separate coding in the random part now yields interpretable insights.

First, the significant random slope effect for offliners ( $\hat{\sigma}_{u1}^2$ ) means that there is variation across interviewers in their success at recruiting offliners. For onliners, we also find varying interviewer effects ( $\hat{\sigma}_{u2}^2$ ). Interestingly, the size of the interviewer variance for offliners is considerably smaller than that of the interviewer variance for onliners ( $\hat{\sigma}_{u1}^2 = 0.49$  vs.  $\hat{\sigma}_{u2}^2 = 1.71$ ). Thus, there is much less variation between interviewers when recruiting offliners compared to onliners (addressing research question 2b). In addition, this difference in the size of the interviewer effects is significantly different for onliners and offliners (Wald  $\chi_{(1)}^2 = 8.38$ ,  $p < 0.01$ ). Finally, the significant positive covariance of the two random slope coefficients ( $\hat{\sigma}_{u1,u2}$ ) indicates that interviewers who are

5. In this case, the negative covariance implies that interviewers who do better than average in recruiting onliners tend to have a below-average difference in success rates for recruiting offliners vs. onliners. It is hard to draw meaningful implications from this finding, however.



	Model 1		Model 2		Model 3	
	Contrast coding		Contrast and separate coding		Cross-level interactions	
	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.
Fixed effects						
Sample unit characteristics						
Being offline	−0.89***	0.18	−0.89***	0.18	−3.19***	0.86
Interviewer characteristics						
Tailoring to adapt to sample units	0.08	0.18	0.08	0.18	−0.37	0.24
Being offline × tailor to adapt to sample units					0.68**	0.25
Intercept	0.41	3.16	0.41	3.16	1.84	3.17
Random effects						
$\hat{\sigma}_{y0}^2$	1.70***	0.37				
$\hat{\sigma}_{u1}$	0.68***	0.33	0.49***	0.22	0.42***	0.20
$\hat{\sigma}_{y0,u1}$	−0.94**	0.32				
$\hat{\sigma}_{u2}^2$			1.71***	0.38	1.66***	0.36

Continued

**Table 2.** *Continued*

	Model 1 Contrast coding		Model 2 Contrast and separate coding		Model 3 Cross-level interactions	
	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.	$\hat{\beta}$	Std. err.
$\hat{\sigma}_{u1,u2}$			0.76***	0.19	0.79***	0.19
Deviance	2,508.48		2,508.48		2,508.36	
Number of sample units	3,751		3,751		3,751	
Number of interviewers	214		214		214	

NOTES.—Three interviewers failed to respond to all questions used in these models. All models control for the sample unit characteristics age, age squared, gender, household size, educational level, occupational status, Internet usage, media consumption, voting behavior, and recruitment year. Furthermore, all models control for the interviewer’s characteristics age, age squared, gender, educational level, an expectation of overall response rate, and adaptation to sample units. The full model is presented in supporting online materials, [appendix A, table A.3](#). The statistical test of the variances was estimated in line with [Molenberghs and Verbeke \(2007\)](#) and [Snijders and Bosker \(2011, pp. 98–9\)](#).  $\hat{\beta}$  = unstandardized beta coefficients; Std. err. = standard errors; \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

good at gaining responses from onliners are also good at gaining responses from offliners and vice versa.

### 6.3 Cross-Level Interactions

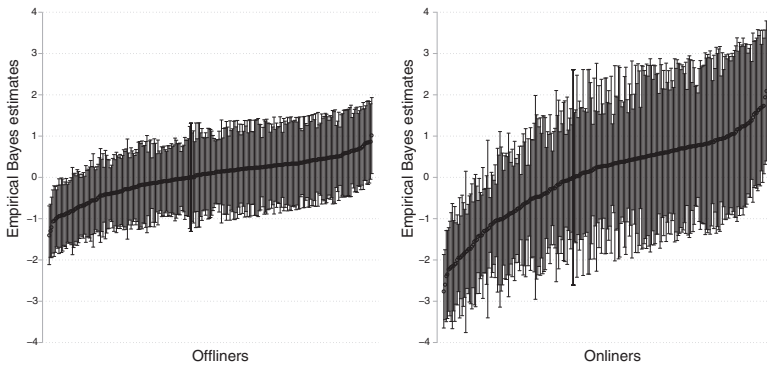
In the third model in [table 2](#), we extend our analysis by estimating cross-level interactions (estimation equivalent to (6) and addressing research question 2c) to find interviewer characteristics that explain the interviewer effects. Five cross-level interactions between interviewer characteristics and respondent characteristics were tested (reported interviewing behavior, work experience, deviating from standardized interviewing protocol, online panel participation rate expectations, and abilities to explain online participation to offliners; details on indicators used can be found in [appendix A, table A.2](#) and [appendix C in the supplementary data online](#)). Of these, only the interaction between being offline and interviewers alleging that they deviate from standardized interviewing protocols to tailor to the respondents' needs is significant.

The main effect for interviewers' tailoring is not significant, whereas the main effect for the respondent characteristic "being offline" and participation in the online panel shows a significant and negative association. The cross-level interaction between being offline and interviewers tailoring as well as the main effect of being offline are significant while the main effect for interviewers tailoring is not significant. The positive interaction effect means that tailoring is more effective when recruiting offliners than when recruiting onliners.

### 6.4 Identifying Underperforming Interviewers

So far, we have demonstrated that separate coding in the random part of the model allows us to identify interviewer effects separately for onliners and offliners and how the ensuing estimation of cross-level interactions may explain the interviewer effects found. In a last step, we would like to illustrate how practitioners might subsequently identify underperforming interviewers who may need to be replaced or receive additional training.

For this purpose, we plot the empirical Bayes estimates of the random slope per interviewer separately for the onliners and offliners (see [figure 1](#)). Each estimate and its respective confidence interval denotes an interviewer. Interviewers are sorted by their estimated empirical Bayes mean from left (worst performing) to right (best performing). When selecting interviewers for replacement or retraining, the survey practitioner would start at the left with the worst-performing interviewers and select as many interviewers as their budget or other constraints permit. Alternatively, if the graph displays a point of steep increase (our example does not), this may also be a suitable cut-off for interviewer selection.



**Figure 1. Empirical Bayes Estimates of the Random Slope per Interviewer, Separately for Onliners and Offliners.**

## 7. DISCUSSION

Despite an abundance of studies on interviewer effects on survey participation (see [West and Blom 2017](#)), the literature is sparse on how face-to-face interviewers differentially affect specific groups of sample units. This is surprising because differential interviewer effects on the recruitment of different groups of gross sample units may lead to imbalances in the net sample. Acknowledging the importance of this issue, this paper proposes an alternative parametrization of the random components in multilevel models, so-called separate coding, that enables the detection of differential interviewer effects for specific sample groups. In addition, cross-level interactions in such models yield insight into which interviewer characteristics differentially affect which sample groups. Finally, we show how survey practitioners may use the model results for fieldwork optimization and how to identify those interviewers that are underperforming in recruiting a certain sample group.

To exemplify this alternative parameterization strategy, this article investigates interviewer effects on the survey participation in a face-to-face recruited online panel that provides equipment and support to previously offline households (offliners) to enable their participation in the online panel. Previous analyses have shown that low response rates among offliners are cause for concern ([Blom et al. 2017](#)).

Our analyses find significant variance at the interviewer level, that is, the face-to-face interviewers affect participation in the online panel. A multilevel logistic regression with separate coding in the random slopes subsequently reveals small interviewer variances for offliners and much larger interviewer variances for onliners. This means that we find different interviewer effects for offliners and onliners (research question 2a). In fact, the interviewer effect for recruiting offliners is approximately three times smaller than for recruiting onliners (research question 2b). However, we find

a significant cross-level interaction for offliners: Interviewers who say that they tailor their approach to the sample person in front of them are better at eliciting participation from offline sample units (research question 2c). Finally, a significant positive covariance suggests that interviewers who are good at gaining response from onlineers are also good in gaining response from offliners and vice versa.

In practice, there are two ways of interpreting these results: If the glass is half empty, changes to the fieldwork processes aimed at improving the performance of underachieving interviewers (e.g., by specifically training them in their approach to offliners) are unlikely to have much impact on the response rates of offliners. Furthermore, the relatively large interviewer variances for onlineers suggest scope for improving interviewer performance for this group. Such an intervention may increase overall response rates, yet it may also further increase differences in response rates between offliners and onlineers. In contrast, if we consider the glass half full, participation among offliners might increase if interviewers are better trained in to tailor their approach, particularly speed and dialect of their speech, to liken the sample units in front of them. According to our models, such tailoring would be especially successful for offliners. However, the overall scope for response rate increases among offliners is likely small.

## 8. LIMITATIONS

The data available to us for this study stem from the GIP survey operations, that is, they were not collected for this study specifically and thus come with limitations. For example, we are unable to perfectly disentangle area effects from interviewer effects. Therefore, future research may gain from an interpenetrated survey design, as it is often implemented in telephone surveys (e.g., see [Lipps 2019](#)). Unfortunately, in the context of face-to-face surveys, most researchers opt for noninterpenetrated designs because of the practical complexities of and high fieldwork costs associated with the random allocation of areas and interviewers.

Due to a lack of validation data, we are also unable to disentangle interviewer effects on measurement from those stemming from the recruitment process (see [West et al. 2018a](#) and [West and Olson 2010](#) for examples). Future research combining such a decomposition with our parametrization of the random slopes will likely be very valuable to the field of interviewer research.

In addition, there is scope for extending our research without a need for expensive new data collection. For example, we have not yet fully exploited the potential offered by the sequential mixed-mode design of the GIP. Future research may investigate whether the interviewer variances found during the GIP recruitment process remain detectable in later online survey waves, that is, whether recruitment interviewers have a lasting effect on the online panel. In

addition, future research may benefit from Bayesian approaches for the variance decomposition to achieve more robust estimates for the smaller offliner sample group (e.g., see [West and Elliott 2014](#)).

We believe that the alternative parametrization of the multilevel model presented in this paper provides a valuable addition to the interviewer effects literature and, importantly, to survey practitioners' toolkit for optimizing fieldwork. We consciously chose to limit our paper to this specific purpose. However, we are aware that our approach generates curiosity regarding the potential biases introduced by interviewers and how using contrast coding to identify underperforming interviewers for specific sample groups may alleviate or aggravate such biases. We hope that our approach will be used as a steppingstone to evaluating such biases and welcome respective extensions by other researchers.

Our approach demands relatively detailed data not only on the interviewers but also on the sample units. We were able to use the two-stage recruitment process of the GIP to feed our models with sample unit information. Most surveys will need to rely on their sampling frame data, which in many countries is very limited. Therefore, we expect that survey practitioners in countries with rich sampling frames, such as the Nordic countries in Europe, might find the insights presented in this paper more practically useful than researchers who need to resort to more limited frames like the area probability sampling frames.

## 9. CONCLUSION

We encourage survey practitioners to test our general recommendations with experiments when fielding interviewer-assisted surveys. In addition, survey methodologists may look into differential interviewer effects in other surveys where they find specific sample groups with very different response rate levels. For example, in a study of adult verbal, mathematical, and analytical competencies, [Helmschrott and Martin \(2014\)](#) found much higher nonresponse rates among lower educated respondents. In such a context, insights into whether interviewers differ in gaining responses from lower educated respondents and potential correlates of such differential interviewer effects might inform interviewer–respondent matching protocols and enable addressing the sample imbalances (e.g., see [West et al. 2020](#)).

During the review process of our manuscript, we were pointed toward an exciting possibility of adopting our approach in the context of adaptive survey designs. In an adaptive design context, the limitation of a need for detailed sampling frame data upfront might be bypassed by collecting more information on the sample units throughout the phases of adaptive data collection. Once sufficient data are available, our approach may be used to identify sample units that may benefit from new fieldwork procedures, for example, with specifically trained interviewers or even interviewer-free (for more information on adaptive

design, see Schouten et al. 2013). In particular, in the context of sequential mixed-mode surveys with an initial interviewer-mediated mode, the adaptation toward a self-completion approach would be interesting to explore. To us, adaptive designs seem like a potentially fruitful avenue that may allow an integration of our approach into highly professional fieldwork optimization strategies.

For the future, we hope that the proposed parameterization strategy for random slopes in multilevel models contribute in two ways: We encourage survey statisticians to further develop this line of research by, for example, integrating Bayesian approaches, extending the modeling to the decomposition of effects on measurement and nonresponse, and integrating our approach into adaptive designs. More importantly, however, we encourage survey practitioners to apply our approach to survey settings where differential recruitment success of interviewers is to be expected. A practical guide to the interpretation of estimates and possible fieldwork actions related to these can be found in this paper.

## Supplementary Material

Supplementary materials are available online at [academic.oup.com/jssam](https://academic.oup.com/jssam/article/1/2/1/249/6686757).

## REFERENCES

- Ackermann-Piek, D., Korbmacher, J. M., and Krieger U. (2020), "Explaining Interviewer Effects on Survey Unit Nonresponse: A Cross-Survey Analysis," in *Interviewer Effects from a Total Survey Error Perspective*, eds. K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, and B. T. West, New York: Chapman and Hall/CRC, pp. 193–206.
- Beullens, K., Loosveldt, G., and Vandenplas C. (2019), "Interviewer Effects among Older Respondents in the European Social Survey," *International Journal of Public Opinion Research*, 31, 609–625.
- Blom, A. G. (2012), "Explaining Cross-Country Differences in Survey Contact Rates: Application of Decomposition Methods," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175, 217–242.
- Blom, A. G., and Korbmacher, J. M. (2013), "Measuring Interviewer Characteristics Pertinent to Social Surveys: A Conceptual Framework," *Survey Methods: Insights from the Field*. Available at <http://surveyinsights.org/?p=817>.
- Blom, A. G., Gathmann, C., and Krieger, U. (2015), "Setting up an Online Panel Representative of the General Population: The German Internet Panel," *Field Methods*, 27, 391–408.
- Blom, A. G., de Leeuw, E., and Hox, J. J. (2011), "Interviewer Effects on Nonresponse in the European Social Survey," *Journal of Official Statistics*, 27, 359–377.
- Blom, A. G., Herzing, J. M. E., Comesse, C., Sakshaug, J. W., Krieger, U., and Bossert, D. (2017), "Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence from the German Internet Panel," *Social Science Computer Review*, 35, 498–520.
- Brunton-Smith, I., Sturgis, P., and Williams, J. (2012), "Is Success in Obtaining Contact and Cooperation Correlated with the Magnitude of Interviewer Variance?," *Public Opinion Quarterly*, 76, 265–286.

- Cornesse, C., and Schaurer, I. (2021), "The Long-Term Impact of Different Offline Population Inclusion Strategies in Probability-Based Online Panels: Evidence from the German Internet Panel and the GESIS Panel," *Social Science Computer Review*, 39, 687–704.
- Couper, M. P., and Groves, R. M. (1992), "The Role of the Interviewer in Survey Participation," *Survey Methodology*, 18, 263–278.
- Durrant, G. B., and Steele, F. (2009), "Multilevel Modelling of Refusal and Non-Contact in House- Hold Surveys: Evidence from Six UK Government Surveys," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 361–381.
- Durrant, G. B., Groves, R. M., Staetsky, L., and Steele, F. (2010), "Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys," *Public Opinion Quarterly*, 74, 1–36.
- Eckman, S. (2016), "Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias?," *Social Science Computer Review*, 34, 41–58.
- Groves, R. M. (2006), "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, 70, 646–675.
- Groves, R. M., and McGonagle, K. (2001), "A Theory-Guided Interviewer Training Protocol Regarding Survey Participation," *Journal of Official Statistics*, 17, 249–266.
- Heerwegh, D. (2009), "Mode Differences between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects," *International Journal of Public Opinion Research*, 21, 111–121.
- Helmschrott, S., and Martin, S. (2014), "Nonresponse in PIAAC Germany," *Methods, Data, Analyses*, 8, 243–266.
- Herzing, J. M. E., and Blom, A. G. (2019), "The Influence of a Person's Digital Affinity on Unit Nonresponse and Attrition in an Online Panel," *Social Science Computer Review*, 37, 404–424.
- Hox, J. J. (1994), "Hierarchical Regression Models for Interviewer and Respondent Effects," *Sociological Methods & Research*, 22, 300–318.
- . (2010), *Multilevel Analysis: Techniques and Applications*, New York and Hove: Routledge.
- Hox, J. J., and de Leeuw, E. D. (2002), "The Influence of Interviewers' Attitude and Behaviour on Household Survey Nonresponse: An International Comparison," in *Survey Nonresponse*, eds. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, New York, NY: Wiley, pp. 103–120.
- Jäckle, A., Lynn, P., Sinibaldi, J., and Tipping, S. (2012), "The Effect of Interviewer Experience, Attitudes, Personality and Skills on Respondent Co-Operation with Face-To-Face Surveys," *Survey Research Methods*, 7, 1–15.
- Jones, K. (2013), *Developing Multilevel Models for Analysing Contextuality, Heterogeneity and Change Using Mlwin 2.2* (Vol. 1), Bristol: Centre for Multilevel Modelling.
- Lipps, O. (2019), "Interviewer Effects on Cooperation during Initial and Refusal Conversion Fieldwork Phases in Telephone Panel Surveys," *Field Methods*, 31, 375–393.
- Lipps, O., and Pollien, A. (2011), "Effects of Interviewer Experience on Components of Nonresponse in the European Social Survey," *Field Methods*, 23, 156–172.
- Little, R. J. A. (1988), "Missing-Data Adjustments in Large Surveys," *Journal of Business & Economic Statistics*, 6, 287–296.
- Loosveldt, G., and Wuyts, C. (2020), "A Comparison of Different Approaches to Examining Whether Interviewer Effects Tend to Vary across Different Subgroups of Respondents," in *Interviewer Effects from a Total Survey Error Perspective*, eds. K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, and B. T. West, Boca Raton: CRC Press, pp. 311–322.
- Loosveldt, G., and Beullens, K. (2014), "A Procedure to Assess Interviewer Effects on Nonresponse Bias," *SAGE Open*, 1–12. doi: 10.1177/2158244014526211.
- Molenberghs, G., and Verbeke, G. (2007), "Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space," *The American Statistician*, 61, 22–27.
- Morris, T. P., White, I. R., and Royston, P. (2014), "Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws," *BMC Medical Research Methodology*, 14, 1–13.
- Rabe-Hesketh, S., and Skrondal, A. (2008), *Multilevel and Longitudinal Modeling Using Stata*, College Station: Stata Press.



- Revilla, M., Cornilleau, A., Cousteaux, A.-S., Legleye, S., and de Pedraza, P. (2016), "What is the Gain in a Probability-Based Online Panel of Providing Internet Access to Sampling Units Who Previously Had No Access?," *Social Science Computer Review*, 34, 479–496.
- Schouten, B., Calinescu, M. and Luiten, A. (2013), "Optimizing Quality of Response through Adaptive Survey Designs," *Survey Methodology*, 39, 29–58.
- Snijders, T. A. B., and Bosker, R. J. (2011), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, London and Thousand Oaks: Sage.
- Steele, F., and Durrant, G. B. (2011), "Alternative Approaches to Multilevel Modelling of Survey Non-Contact and Refusal," *International Statistical Review*, 79, 70–91.
- Verbeke, G., and Molenberghs, G. (2000), "Inference for the Marginal Model," in *Linear Mixed Models for Longitudinal Data*, eds. G. Verbeke and G. Molenberghs, New York: Springer, pp. 55–76.
- West, B. T. (2020), "Designing Studies for Comparing Interviewer Variance in Two Groups of Survey Interviewers," in *Interviewer Effects from a Total Survey Error Perspective*, eds. K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, and B. T. West, New York: Chapman and Hall/CRC, pp. 323–333.
- West, B. T., Ong, A. R., Conrad, F. G., Schober, M. F., Larsen, K., and Hupp, A. (2022), "Interviewer Effects in Live Video and Recorded-Video Interviewing," *Journal of Survey Statistics and Methodology*, 10, 317–336.
- West, B. T., and Blom, A. G. (2017), "Explaining Interviewer Effects: A Research Synthesis," *Journal of Survey Statistics and Methodology*, 5, 175–211.
- West, B. T., and Olson, K. (2010), "How Much of Interviewer Variance Is Really Nonresponse Error Variance?," *Public Opinion Quarterly*, 74, 1004–1026.
- West, B. T., and Elliott, M. R. (2014), "Frequentist and Bayesian Approaches for Comparing Interviewer Variance Components in Two Groups of Survey Interviewers," *Survey Methodology*, 40, 163–188.
- West, B. T., Conrad, F. G., Kreuter, F., and Mittereder, F. (2018a), "Nonresponse and Measurement Error Variance among Interviewers in Standardized and Conversational Interviewing," *Journal of Survey Statistics and Methodology*, 6, 335–359.
- . (2018b), "Can Conversational Interviewing Improve Survey Response Quality without Increasing Interviewer Effects?," *Journal of the Royal Statistical Society (Series A)*, 181, 181–203.
- West, B. T., Kreuter, F., and Jaenichen, U. (2013), "'Interviewer' Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?," *Journal of Official Statistics*, 29, 277–297.
- West, B. T., Elliott, M. R., Mneimneh, Z., Wagner, J., Peytchev, A., and Trappmann, M. (2020), "An Examination of an Interviewer-Respondent Matching Protocol in a Longitudinal CATI Study," *Journal of Survey Statistics and Methodology*, 8, 304–324.