Disentangling Interviewer and Area Effects in Large-Scale Educational Assessments using Cross-Classified Multilevel Item Response Models

Theresa Rohm[1, 2], Claus H. Carstensen[2], Luise Fischer[1, 2], and Timo Gnambs[1, 3]

[1] Leibniz Institute for Educational Trajectories

[2] University of Bamberg

[3] Johannes Kepler University Linz

Author Note

Theresa Rohm, Leibniz Institute for Educational Trajectories and University of Bamberg, Germany; Claus H. Carstensen, University of Bamberg, Germany; Luise Fischer, Leibniz Institute for Educational Trajectories and University of Bamberg, Germany; Timo Gnambs, Leibniz Institute for Educational Trajectories, Germany, and Johannes Kepler University Linz, Austria.

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6 – Adults, doi:10.5157/NEPS:SC6:3.0.1. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Correspondence concerning this article should be addressed to Theresa Rohm, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, Email: theresa.rohm@lifbi.de

Accepted for publication in the *Journal of Survey Statistics and Methodology*.

Abstract

In large-scale educational assessments, interviewers should ensure standardized settings for all participants. However, in practice many interviewers do not strictly adhere to standardized field protocols. Therefore, systematic interviewer effects for the measurement of mathematical competence were examined in a representative sample of N = 5,139 German adults. To account for interviewers working in specific geographical regions, interviewer and area effects were disentangled using cross-classified multilevel item response models. These analyses showed that interviewer behavior distorted competence measurements, whereas regional effects were negligible. On a more general note, it is demonstrated how to identify conspicuous interviewer behavior with Bayesian multilevel models.

Keywords: Administration effects, Competence measurement, Interviewer effects, Large-scale assessment, Multilevel item response theory.

Introduction

Interviewer behavior is an essential factor in large-scale educational assessments to guarantee valid measurements of, for example, cognitive abilities, motivations, or attitudes (Moss et al. 2006). By adhering to standardized field protocols, interviewers need to accomplish a variety of tasks such as creating comparable settings that avoid unnecessary disruptions or providing similar assistance to all participants that does not give an undue advantage to some respondents. Typically, not all interviewers are equally capable; depending on their abilities or motivations some interviewers might be more likely to succeed in creating standardized assessment conditions than others (Schaeffer et al. 2010; Turner et al. 2014; West and Olson 2010; West et al. 2013). If specific interviewer behavior affects the responses of some participants, responses from different respondents being assessed by the same interviewer are likely correlated and, thus, exhibit a systematic interviewer-specific variance. This variance might even depend on specific interviewer characteristics (e.g., age or experience) or interactions between interviewer and respondent characteristics. Consequently, a test taker's responses not only reflect the construct of interest (i.e., attitudes, cognitive abilities) but also context effects introduced by non-standardized assessment conditions. Ignoring these dependencies in the analysis of respondent data risks underestimating standard errors and, in turn, overestimating the statistical significance of effects (Durrant et al. 2010; Finch and Bolin 2017).

Because interviewers often work in a specific geographical region, interviewer effects can be confounded with regional characteristics (O'Muircheartaigh and Campanelli 1998). To deal with the possible relatedness of respondents being assessed by the same interviewer and of respondents living in the same sampling area, the present study adopts cross-classified multilevel models to disentangle both sources of variance. In this way, dependencies introduced by interviewer behavior and geographical areas are distinguished by estimating separate random effect structures (Maas and Hox 2004). We demonstrate cross-classified

multilevel modeling in a German large-scale assessment of mathematical competences and evaluate the impact of interviewers on competence measurement.

Interviewer and Area Effects in Large-Scale Assessments

Domain-specific competences such as mathematical or reading competence represent central factors for successful performance in many educational and professional situations (Hartig et al. 2008). They explain educational trajectories, occupational choices, and even differences in wages (Heckman et al. 2006) and, thus, determine the social and economic success of individuals. Moreover, from the perspective of cross-country comparisons, enhanced cognitive skills improve economic well-being of nations (Hanushek and Woessmann 2008). Therefore, various large-scale assessments such as the *Programme for International Student Assessment* (PISA), the *Trends in International Mathematics and Science Study* (TIMMS), the *Progress in International Reading Literacy Study* (PIRLS), or the *Programme for the International Assessment of Adult Competencies* (PIAAC) have been initiated to identify determinants of skill inequality and provide policy makers recommendations for political action. The study of competences requires standardized measurements that allow for the estimation of reliable competence scores (Pohl and Carstensen 2013). Importantly, respective test scores should only reflect individual differences in the measured competence and not situational influences or context effects from, for example, different assessment modes (e.g., computer versus paper; cf. Wang et al. 2007), distractive environments (e.g., disturbance by other test takers or media devices; cf. Shelton et al. 2009), or different forms of assistance (e.g., lengthy versus limited test instructions). In this regard, interviewers are assigned an essential role. They are responsible for the implementation of standardized assessment settings for all participants and, thus, should give each test taker equal opportunities to achieve good test scores.

Interviewer Effects

Interviewers can affect the quality of the obtained data through the contact with possible respondents (nonresponse error) and the actual process of interviewing (interviewer bias). Nonresponse error is produced because interviewers influence the propensity of the respondents to participate in the survey (Schaeffer et al. 2010; West and Olson 2010; West et al. 2013; Vassallo et al. 2015). In contrast, interviewer bias is introduced during the administration of the questionnaire or test. Various directly observable interviewer characteristics such as the interviewers' age, gender, or ethnicity as well as unobservable characteristics (e.g., experiences, stereotypes about the respondent, attitudes toward the surveyed topic, expectations about item difficulty) can exert nonnegligible effects on survey responses (Brunton-Smith et al. 2012; Groves 1989; Hox 1994; O'Muirchertaigh and Campanelli 1998; Rosenthal 1967, 2002; Tourangeau and Yan 2007). For example, a well-known systematic influence on survey results are interpersonal expectancy effects (Rosenthal 1994). Interviews are a social process: not only respondents provide information to the interviewer, but also interviewers provide information to the respondents. If interviewers hold certain beliefs about the topic of a survey, they might unintentionally communicate subtle hints (e.g., by body language, tone of voice) to which respondents might react. In survey research, interviewer effects have sometimes been found to be small, often explaining less than 10 percent of variance in nationally representative household surveys (e.g., Brunton-Smith et al. 2016; Groves 1989). Rarely, cross-country studies show intra-interviewer variance approaching 20 percent (e.g., Beullens and Loosveldt 2014, 2016). However, even small effects can have an undue impact on the quality of the obtained data, particularly when each interviewer surveys many respondents (Collins 1980; Hox et al. 1991; Kish 1965; Schaeffer et al. 2010).

Differences in interviewer behavior can also systematically bias the assessment of competences (Rosenthal 1994, 2002). Although great effort is invested into standardizing

large-scale assessments, for example, with the help of administration manuals and mandatory interviewer trainings, empirical investigations on the effectiveness of these efforts is rather limited. One exception is an analysis of interviewer effects within institutional contexts (classroom setting) of student educational assessments (Lüdtke et al. 2007). These authors found negligible interviewer effects in the 2002 PISA assessment of mathematical competence explaining less than one percent of variance. Furthermore, neither interviewer characteristics (e.g., gender, experience) nor interactions between respondents' and interviewers' gender yielded an effect on the observed achievement scores. So far, little is known about interviewer effects on competence measurement in non-institutional individual settings. Because of the less standardized assessment situation in the respondents' private homes, differences in interviewer behavior might have stronger effects on competence measurements.

Area Effects

Another source of imprecision in the estimation of respondents' proficiency is variance introduced through the sampling of respondents through regional clusters. In complex sampling designs, respondents are selected from a population using multistage cluster sampling. Thereby, the responses of survey participants belonging to the same area cluster can be correlated. The homogenizing effect of sampling points is also termed "spatial homogeneity" (Schnell and Kreuter 2005). It results from similar sociodemographic characteristics of respondents who live in the same area (Gabler and Lahiri 2009; Schnell and Kreuter 2002), as well as socio-economic and cultural characteristics, accessibility or factors of urbanicity (Haunberger 2010). For example, within a regional cluster income, age, and ethnicity of the respondents are likely to be more similar than across different clusters (Lee et al. 1989); consequently, measured attitudes, proficiencies, and behaviors related to these characteristics are likely to be correlated with the regional clustering.

In face-to-face surveys, interviewers are often assigned to respondents based on spatial proximity. However, when each interviewer works in a specific region, effects of interviewers and areas can be confounded. This confounding could be minimized with the use of an interpenetrated design (Hox 1994; Mahalanobis 1946), where interviewers are assigned at random to respondents, living in different areas. Consequently, explanatory variables on the interviewer and area level can be assumed to be no longer correlated. However, this is often rather impractical for national surveys because this design is associated with high travel expenses for interviewers. In contrast, partially or limited interpenetrated designs allow to empirically disentangle interviewer and area effects although interviewer and area clusters do overlap to some extent. As a requirement for this limited interpenetrated design, some interviewers work in more than one area and areas are visited by more than one interviewer. For example, a recent simulation study (Vassallo et al. 2017) on cross-classified multilevel logistic models predicting survey non-response suggests that already three areas per interviewer can be sufficient interviewer dispersion across areas, resulting in good precision of survey estimates. Particularly, the random variance structure can be severely biased when interviewers work in only one area, whereas intercept estimates seem to be less affected by restrictive interviewer allocation schemes.

Therefore, empirical analyses of interviewer effects need to account for possible clustering of interviewers within specific areas (Brunton-Smith et al. 2012; Durrant et al. 2010; Turner et al. 2014). In survey research, joint estimations of interviewer and area effects typically found that interviewers made a higher contribution to the homogenizing effect in survey estimates as compared to sampling point clusters (Hansen et al. 1961; O'Muircheartaigh and Campanelli 1999; Schnell and Kreuter 2002). These studies were similar in trying to randomly allocate respondent addresses to interviewers within geographical pools or districts. Main differences were the purpose of the study (accuracy of U.S. census data vs. refusal and non-contact in the British Household Panel Study vs. a

design-effects study), the strategy of random allocation of interviewers to areas (e.g., the amount of interviewers allocated to respondents within and across areas) and the statistical method used to separate the interviewer and area effects (F-test vs. multilevel cross-classified models vs. three-level models). However, in large-scale educational assessments of adult competencies, confounded interviewer and area effects have rarely been investigated.

Identification of Interviewer and Area Effects

Multilevel modelling is useful to separate construct variance from context effects. If substantial interviewer or area effects occur, individual observations are not completely independent. These dependencies can be acknowledged in the modeled error structure by specifying different random effects (Maas and Hox 2004). Multilevel cross-classified models allow for more than one effect of nesting to occur at the same level (Raudenbush 1993; Rasbash and Goldstein 1994; Goldstein 2011). Hence, they can alleviate the problem of confounded effects that occur from interviewer nesting and spatial clustering (Durrant et al. 2010; Hox and DeLeeuw 1994; O'Muircheartaigh and Campanelli 1998). Especially when the implementation of a completely interpenetrated design (respondents are randomly assigned to interviewers, independent of any regional allocation) is not feasible, multilevel modelling approaches are beneficial to obtain unbiased estimates from partially interpenetrated designs.

To investigate cross-classified interviewer and area effects in competence measurement, we adopt a multilevel structural equation modelling (SEM) framework where the measurement part is specified as a two-parameter item response theory (IRT) model as $y_i^* = \Lambda \cdot \theta_i + \varepsilon_i$ (Kamata and Vaughn 2011). Here, y_i^* represents the vector of J unobserved latent response variables for respondent $i \in 1 \ldots I$ that gives rise to the observed dichotomous responses y_i such that for item j $y_{ij} = 1$ if $y_{ij}^* \geq \tau_j$ and $y_{ij} = 0$ if $y_{ij}^* < \tau_j$. The latent variable θ_i is a vector of K factor scores representing the measured ability; in case of a unidimensional model, $K = 1$. Finally, Λ is an I x K matrix of discrimination parameters and ε are the I zero mean normally distributed residuals.

The structural model part allows for varying intercepts and regression coefficients across *C* interviewers and *G* area clusters (for further details see Kamata and Vaughn 2011; Kaplan 2014). This can be expressed as

$$\theta_{icg} = \alpha_{cg} + \Gamma_{cg} x_{icg} + u_{icg} \text{ with } c = 1, 2, \ldots, C \text{ and } g = 1, 2, \ldots, G, \quad (1)$$

where the latent factor θ_{icg} for respondent *i* nested in interviewer *c* and area *g* is regressed on individual-level covariates x_{icg}. The intercept α_{cg} and the slopes Γ_{cg} are allowed to vary across interviewers and areas as a function of between-interviewer variables w_c and between-area variables w_g:

$$\alpha_{cg} = \alpha_{00} + A_c w_c + A_g w_g + \varepsilon_c + \varepsilon_g \quad (2)$$

$$\Gamma_{cg} = \Gamma_{00} + \Gamma_c w_c + \Gamma_g w_g + \xi_c + \xi_g . \quad (3)$$

The residual u_{icg} is assumed to be normally distributed with zero mean and variance $Var(u_{icg}) = \sigma^2_u$, whereas the residuals for the interviewer, ε_c and ξ_c, and area cluster, ε_g and ξ_g, are each multivariate normally distributed with zero means and variance-covariance structures $\Sigma = \begin{pmatrix} \sigma^2_\varepsilon & \sigma_{\varepsilon\xi} \\ \sigma_{\varepsilon\xi} & \sigma^2_\xi \end{pmatrix}$. The three residual structures σ^2_u, Σ_c, and Σ_g are usually assumed to be independent. As the interviewer-to-area distribution is not random due to the design of the analysed study, Σ_c and Σ_g might be correlated. Consequently, even though theoretically assumed, the model cannot reveal interviewer-by-region interaction effects. When the model is presented as an unconditional cross-classified model without predictor variables, (1) and (2) reduce to the mixed-effects formulation

$$\theta_{icg} = \alpha_{00} + \varepsilon_c + \varepsilon_g + u_{icg} . \quad (4)$$

Here, the achievement of respondent *i* equals the sum of the grand-mean achievement of all respondents α_{00}, the random effect ε_c introduced by interviewer *c*, the random effect ε_g of the region *g*, and a random respondent effect u_{icg}.

Cross-classified multilevel models can be estimated in a Bayesian framework with a Markov Chain Monte Carlo (MCMC) algorithm. Parameter estimates are obtained from posterior distributions using a Gibbs-sampler that are generated by repeated sampling from conditional distributions based upon observed data given prior information about the parameters. Thus, the uncertainty about parameter estimates is reflected in the posterior distribution. This allows for the calculation of point estimates (posterior mean) and posterior credibility intervals, which do not rely on a normal approximation of the posterior distribution (Van den Noortgate et al. 2003). Nevertheless, using non-informative priors leads to results that are asymptotically equivalent to respective maximum likelihood estimates (Muthén and Asparouhov 2016). The Bayesian method using MCMC estimation has several advantages: For one, complex multilevel models can be fitted to the data that might not be estimable using likelihood-based frequentist methods (Finch and Bollin 2017). Furthermore, the method is helpful for non-continuous (binary) item responses with missing values and unbalanced designs. MCMC-based Bayesian models for binary responses have been examined by Fox and Glas (2001) or Goldstein and Browne (2005); respective MCMC-based Bayesian approaches for continuous and ordinal responses are described in Lee and Song (2004). The flexibility of MCMC-based Bayesian methods for model fitting is especially beneficial for the structural model part of multilevel models, as it does not rely on asymptotic theory, presents posterior distributions for random effects, and results in more accurate parameter estimates (Muthén and Asparouhov 2012; for an application of a Bayesian multilevel SEM, see Kaplan 2014).

Cross-classified multilevel latent variable models are not often presented in large-scale educational assessments. There are applications, for example, in the context of school effectiveness research (Fox 2010) and for the measurement of attainment targets of Dutch reading comprehension for students at the end of primary school (Van den Noortgate, De Boeck, and Meulders 2003). In addition, multilevel cross-classified testlet models were explored to analyse the dependency of items from clustering factors, such as testlet and

content areas, as well as person factors (Jiao, Kamata and Xie 2016). Furthermore, there are applications to longitudinal data, where for example students' performance scores are clustered within students and within teachers (Luo and Kwok 2012). In addition, cross-classified structural equation models were examined for a longitudinal measurement of teacher-ratings of U.S. students' aggressive-disruptive behavior (Asparouhov and Muthén 2016).

In the present study, multilevel models with cross-classifications of interviewer and area clusters are presented to account for possibly confounded effects on the measurement of adult mathematic competence. These analyses have two aims: first, we want to identify to what degree competence measurements in large-scale assessments are distorted by interviewer and area effects. For this purpose, interviewer and area residual variance (σ^2_ε and σ^2_ξ) are set in relation to the overall residual variance (σ^2_u) of the latent factor (θ_{icg}). Second, we demonstrate with a hands-on example how to identify interviewers that unduly influence the test results, based on the random effect variance (σ^2_ε) introduced by interviewer *c*.

Present Study

Our methodological goal is to identify interviewer effects on mathematic achievement through multilevel cross-classified analysis using Bayesian MCMC methods. This is very similar to the aim of Lüdtke et al. (2007). Nevertheless, our study differs in classification factors (test administrator and school vs. interviewer and area), study population (school students vs. adults), setting (group testing vs. face-to-face settings), as well as the modeling of the latent construct. While Lüdtke et al. (2007) used manifest mathematic scores at the respondent level that were scaled in advance of the analysis, we incorporate the measurement model directly into our cross-classified multilevel model. As the mathematic construct cannot be assessed directly, it is measured by a set of items reflecting the hypothetical construct. Thereby, the variable of interest cannot be measured perfectly and, in effect, measurement error is present. Using a cross-classified multilevel latent variable model, we account for

measurement error of the latent variable. The measurement part of our model is specified as a two-parameter logistic IRT model, which is a model that is frequently presented in educational and psychological measurement on multilevel IRT modeling (e.g., Fox 2003; Fox and Glas 2001; Skrondal and Rabe-Hesketh 2004). In comparison to the Rasch Model (Rasch 1980), the assumption of equal item discriminations is relaxed to allow that items discriminate unequally among respondents with different abilities.

Method

Sample and Procedure

The participants were part of the National Educational Panel Study (NEPS; Blossfeld et al. 2011) that included a representative sample of German adults (see Hammon et al. 2016, for details on the stratified multistage sampling procedure). Primary sampling units of a two-stage sampling procedure served as area clusters (strata) in the analyses. Respondents were randomly drawn from local registers of residents within each area cluster and a private research institute supervised the distribution of addresses to interviewers. Although respective information was not explicitly provided, we assume that respondents were allocated to interviewers based on proximity of the living addresses.

Originally, 5,245 respondents participated. However, about two percent of the sample was excluded due to an excessive number of missing values on the competence test (n = 24), background information on the respondents or interviewers (n = 81), or failure to match respondent records to an interviewer (n = 1). Thus, the analyses are based on a sample of N = 5,139 respondents (50.9% women) aged between 25 and 72 years (*Mdn* = 52.33, *SD* = 10.96). Nearly half of the sample attained matriculation standard or holds a graduate degree (47.1%). Overall, the respondents lived in 92 area clusters (strata) with an average of *Mdn* = 37.5 (*Min* = 1, *Max* = 360) persons per regional cluster. The respondents were interviewed by 200 different interviewers (40.0% women) that each tested *Mdn* = 21 (*Min* = 1, *Max*= 123) persons (three interviewers interviewed only one respondent). Furthermore, the interviewers

visited *Mdn* = 2 regions (*Min* = 1, *Max* = 8); each region was visited by *Mdn* = 3 interviewers (*Min* = 1, *Max* = 30). The distribution of regions per interviewer (see online supplement) shows that most of the interviewers (*n* =114, 57%) worked in at least two different regions. Nevertheless, a considerable number of interviewers (*n* =86, 43%) worked in only one region. The respondents were tested individually in their private homes by a professional survey institute. The interviewers had the complex task of administering the competence test within a computer-assisted personal interview. Thus, they had to switch from being responsive to the respondent during the completion of the computerized questionnaire to the application of strict rules of standardization during the subsequent paper-based competence test (Fellenberg et al. 2016). Important tasks for the interviewers during the personal interview were to motivate the respondent and to present the items and response options, whereas they had to standardize the competence assessment by minimizing disturbances in the respondents' home environment. Further details on the data collection process and the survey execution are provided on the project website (http://www.neps-data.de).

Instruments

Mathematical competence was measured with a paper-based achievement test including twenty-one items that were specifically constructed for administration in the NEPS. All items were accompanied by multiple choice or short constructed response formats that were dichotomously scored. The construction rationale and development of the test are described by Neumann and colleagues (2013). Following the NEPS framework for mathematical competence, each item belonged to one of four content areas: (1) quantity, (2) space and shape, (3) change and relationships, and (4) data and chance. Thereby, the content areas of the NEPS do not follow the canonical categorization of mathematical disciplines (e.g., geometry, algebra, analysis, probability theory) but refer to four content areas encompassing everyday problems. Mathematic competence in adulthood is characterized by a strong focus on the literacy aspect (e.g., apply mathematical concepts to a variety of contexts)

as compared to younger cohorts (e.g., students at school). Hence, the measured concept is assumed to have high variance among the adult population, with some items covering mathematical issues that are necessary for everyday life and other items being very specific for typical contexts (e.g., relevant for specific careers/occupations). In addition, but not related to the content areas, six cognitive components were required to solve the tasks: (1) mathematical communication, (2) mathematical argumentation, (3) modeling, (4) using representational forms, (5) mathematical problem solving, and (6) technical abilities and skills. These cognitive processes condition the mathematic ability of adults as they need to be activated when solving the respective item. Both dimensional concepts, the four content areas and the six cognitive components, are closely linked to the PISA framework (OECD, 2004; for details see Neumann et al. 2013). Despite the different components specified in the construction rationale, these are not assumed to represent distinct dimensions. Rather, the test is dominated by a single mathematical factor. In-depth psychometric analyses corroborated a unidimensional structure and measurement invariance across several respondent characteristics (see Jordan and Duchhardt 2013). In addition, hierarchical IRT models with random discrimination and threshold effects did not indicate substantial item variances across interviewer or area clusters.

We acknowledged several *respondent characteristics* that might be associated with mathematical competence: age (in years), gender (coded 0 for men and 1 for women), ethnicity (coded 0 for no migration background and 1 otherwise), educational level (with four categories: lower secondary degree or less, secondary education, matriculation standard, graduate degree), employment status (coded 0 as employed and 1 otherwise), and cultural capital (as reflected by the number of books in the household). In addition, the political area size per respondent (measured as number of inhabitants of the respondents' municipality with seven categories ranging from "below 2,000 inhabitants" to "500,000 and more inhabitants") was acknowledged. Considering these individual characteristics in our analyses should

increase the comparability between regional clusters, because geographical areas might differ on key sociodemographic characteristics. Otherwise, differences between interviewers could reflect differences between areas.

Finally, several *interviewer characteristics* were available. Besides gender (coded 0 for men and 1 for women), the interviewers' age and educational attainment were each measured with three categories using either "less than 50 years", "50-65 years", and "older than 65 years" or "up to lower secondary degree", "secondary education" and "matriculation standard". Work experience as an interviewer, recorded as the general interviewing experience of being employed at the private institute that supervised the assignment of interviewers to the sampled respondents of the NEPS, was indicated on four categories including "up to 2 years", "2-3 years", "4-5 years", and "more than 5 years". Descriptive statistics for these variables are summarized in the online Supplementary Material.

Statistical Analyses

As previous analyses supported a unidimensional scale (Jordan and Duchhardt 2013), a unidimensional two-parametric IRT model (Kamata and Vaughn 2011) was fitted to the mathematical test. Continuous predictors of the latent ability (i.e., respondents' age, number of books in the household, and political area size) were grand-mean centered. Because interviewer effects were expected to be statistically confounded with effects at the area level, cross-classified multilevel models with MCMC and noninformative priors were estimated in *Mplus* 8 (Muthén and Muthén 1998-2017). All variance parameters were estimated using inverse-gamma priors *IG* (-1, 0), while loading and threshold parameters were estimated with normal distribution priors of zero mean and variance of 5, N (0, 5). The prior for the parameters of all first- and second-level covariates was the normal distribution with zero mean and infinity variance, N (0, ∞). A discussion and additional model estimation results on the sensitivity of variance components to the choice of prior can be found in the online Supplementary Material.

All parameter estimates and standard errors are the means and standard deviations of two parallel MCMC chains using a burn-in of half of the minimum 5,000 iterations. Thinning of the chains was applied to reduce autocorrelations (use of every 20^{th} iteration). A convergence criterion of 0.05 was set for each model, indicating that parameter convergence is achieved when the *Potential Scale Reduction* (PSR) values fall below 1.05. Trace plots were used for each parameter to evaluate successful convergence of the estimates. Likewise, autocorrelation function plots were investigated to determine whether the estimated models delivered reliable estimates. For the evaluation of model parameters, the mean of the posterior distribution and the Bayesian 95% credibility interval were used. Posterior predictive checks that compared the predictive distribution to the observed data involved the *Potential Scale Reduction* (PSR) criterion (Gelman and Rubin 1992) for which values below 1.1 indicate convergence (Gelman et al. 2004) and the Kolmogorov-Smirnov test. The latter evaluates the hypothesis that both MCMC chains have an equal distribution, using 100 draws from each of the two chains per parameter.

Finally, the Bayesian residual estimates are used to visualize heterogeneity stemming from interviewer and area clusters, as well as the dependence between units nested within the clusters. To identify exceptional interviewer and area clusters, the posterior standard deviations are used as standard errors for making inferences about the random interviewer and area effects of interest. Random effects are drawn for each cluster based on the posterior distribution of θ_{icg} given the observed data for the cluster. The random effects distribution can thereby be viewed as mirroring the variation of θ_{icg} in the survey population (Skrondal and Rabe-Hesketh 2009). In addition, the posterior standard deviation is used to form confidence intervals for the estimated random intercepts of interviewer- and area-specific measured competence values.

Data Availability and Analyses Syntax

The complete data set analyzed in this study is available at http://www.neps-data.de. Moreover, the analyses syntax used to generate the reported results is provided in an online repository at https://osf.io/fka9x/.

Results

Because respondents were nested in interviewers and geographical areas, mathematical competence was modeled in a cross-classified multilevel IRT framework as outlined above. We estimated a series of increasingly complex models to evaluate potential interviewer effects (see Table 1). The trace and autocorrelation plots for all models indicated sufficient convergence of the parameter estimation. Moreover, after 1,000 iterations the PSR criterion fell below 1.1 for all parameters and the Kolmogorov-Smirnov statistics were not significant (all *p*s > .01). Thus, the models showed appropriate posterior predictive quality for the parameters on the within and between level.

Interviewer and Area Effects

In the first step, we estimated the amount of variance in competence measurement that is attributable to the different interviewers and areas without considering any predictors (i.e., a null model; see Model 1 in Table 1). The impact of clustering on the outcome variable was investigated using intraclass correlation coefficients (ICC) that indicate the proportion of variance attributable to a higher-order cluster (i.e., interviewers, areas) in the total variance. Larger ICCs indicate larger dependencies for interviewer or area clusters and, thus, a greater need for multilevel analyses (Finch and Bollin 2017; Hox 2010). The variance in mathematic achievement between interviewers was much higher than the variance between areas: about 6.6 percent of the observed variance in mathematical competence was attributable to interviewers, whereas only 0.8 percent was attributable to the nesting of respondents in geographical areas.

Table 1. Results of cross-classified multilevel IRT models estimating adult mathematic achievement

	Model 1			**Model 2**			**Model 3**		
	M	*SD*	95% PPI	*M*	*SD*	95% PPI	*M*	*SD*	95% PPI
Fixed effects									
Age				-0.175	0.014	(-0.202, -0.147)	-0.175	0.014	(-0.202, -0.147)
Gender (ref. male)				-0.325	0.012	(-0.348, -0.301)	-0.324	0.012	(-0.348, -0.300)
Migration Background (ref. no)				-0.064	0.013	(-0.089, -0.040)	-0.064	0.013	(-0.090, -0.039)
Educational Attainment (ref. secondary education)									
no degree or lower sec. degree				-0.149	0.015	(-0.178, -0.120)	-0.149	0.015	(-0.179, -0.120)
matriculation standard				0.176	0.014	(0.146, 0.204)	0.175	0.014	(0.146, 0.203)
graduate degree				0.347	0.015	(0.318, 0.376)	0.347	0.015	(0.318, 0.376)
Employment status (ref. employed)				-0.044	0.014	(-0.070, -0.017)	-0.044	0.014	(-0.071, -0.017)
Cultural capital				0.165	0.015	(0.136, 0.194)	0.165	0.015	(0.136, 0.194)
Political Area Size				-0.041	0.019	(-0.078, -0.005)	-0.041	0.019	(-0.079, -0.004)
Interviewer Level Covariates									
Gender (ref. male)							-0.139	0.088	(-0.308, 0.040)
Age (ref. up to 49 years)									
50 to 65 years							-0.091	0.102	(-0.293, 0.109)
older than 65 years							-0.007	0.107	(-0.197, 0.217)
Educational Attainment (ref. lower sec. degree)									
Secondary education							0.150	0.128	(-0.095, 0.401)
Matriculation standard							0.011	0.129	(-0.239, 0.260)
Work experience as interviewer (ref. up to two years)									
2 to 3 years							0.087	0.129	(-0.169, 0.345)
4 to 5 years							0.248	0.126	(-0.016, 0.488)
more than 5 years							0.045	0.126	(-0.202, 0.289)
Variance components of random effects									
Respondents	0.425	0.033	(0.363, 0.492)	0.236	0.020	(0.197, 0.275)	0.244	0.019	(0.209, 0.284)
Interviewers	0.030	0.006	(0.020, 0.045)	0.032	0.006	(0.022, 0.046)	0.033	0.006	(0.023, 0.046)
Areas	0.004	0.003	(0.000, 0.013)	0.001	0.001	(0.000, 0.005)	0.001	0.001	(0.000, 0.005)

Note. Standardized results are presented for fixed effects. *M* = posterior mean. *SD* = posterior standard deviation. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Moreover, the design effect highlights the accuracy of the results in comparison to random sampling; at the same time, it denotes how much larger the sample size must be to obtain the same precision in survey estimates (Schnell and Kreuter 2005). For example, a design effect of 2 is assumed to reduce the effective sample size by half (Schaeffer et al. 2010). In the present study, the design effects for the interviewer and area clusters were 2.60 and 1.44, respectively. Thus, there was substantial interviewer variance, but negligible area effects.

In the second step (see Model 2 in Table 1), the respondent characteristics and the size of the political area the respondents live in were added as fixed effects. This revealed significantly ($p < .05$) worse achievement for women, respondents with migration background, lower education, or a lower socio-economic status, and those without employment. Moreover, test takers living in smaller areas (as measured by political area size) achieved slightly better mathematical competence as compared to people living in strongly populated areas. Although the inclusion of these variables reduced the respondent-specific random variance by nearly a half, the interviewer variance remained unaffected.

Identification of Influential Interviewers

Even though the variance in mathematical competences traceable to interviewer presence was rather high, none of the investigated interviewer characteristics (e.g., gender, age, education, and work experience) was found to be significantly related to the latent competence of the respondents (see Model 3 in Table 1). Furthermore, the interaction of interviewer and respondent gender did not affect mathematical achievement. Thus, socio-demographic differences were unable to identify interviewers with aberrant test administration behaviors. Therefore, we used the interviewer residual terms (second level errors) that were sampled from the posterior distribution of our estimated multilevel model (Model 1) to identify exceptional interviewers. Because these residuals were sample estimates and, therefore, incorporated a level of uncertainty (e.g., they depend on the number of

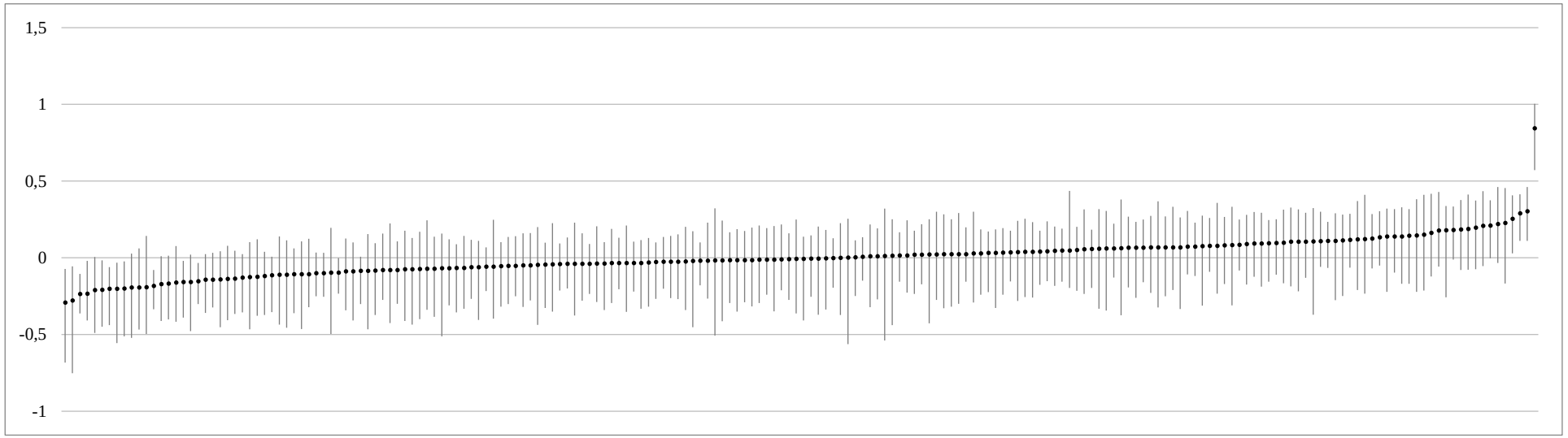
interviewed respondents and on the amount of within- and between-interviewer variation), we ranked the interviewers according to the interviewer residual effects with their 95% probability interval (see Figure 1). Residuals whose posterior probability intervals do not overlap with the general mean indicate interviewers with undue influence on the competence measurement of the respondents. Out of 200 interviewers, 4 had an interval above and 12 had an interval below zero. Hence, their estimated competence intercept deviates from the survey population mean.

To confirm that the results did not depend on the number of regions an interviewer worked in, we refitted Models 1 to 3 to data collected by the 57% of interviewers who worked in at least two different regions. These results did not indicate substantial differences from the findings reported above, as the estimated fixed and random effects remained nearly identical (see Table S12 in the online supplementary material). In addition, interviewer residual effects whose 95% posterior probability interval did not overlap with the general mean in the original estimation, also had significant deviation in their residual effect in these sensitivity analyses.

In multilevel IRT analyses shrinkage to the general mean can be expected for interviewer residuals if a large number of respondents were assigned to an interviewer. Then the posterior mean resembles the intercept of a separate regression for this interviewer. Hence, the identification of exceptional interviewers also depends on the group size (i.e., the number of respondents per interviewer), which is also termed sensitivity of interviewer residuals to group size (Pickery and Loosveldt 2004). However, in our case the interviewer residual was not correlated to the number of completed test administrations, $r = .07$, $p = .30$. Thus, the amount of uncertainty on the interviewer level did not depend on the size of the clusters. In addition, as 3 out of the 200 interviewers only interviewed one respondent, we refitted all models excluding these three cases. The exclusion did not alter the results presented above. Finally, we tested the sensitivity of individual assessment scores to the presence of random interviewer effects, by using posterior means of estimated mathematic competence.

Comparing these estimated values of individual assessments between (1) the model with random interviewer effects and (2) the model ignoring the nested structure (hence, the 2PL model) resulted in a high correlation ($r = .97$, $p = \leq .001$), an average mean deviation in individual competence scores of .00, and a root-mean-squared error of .22. In conclusion, interviewer differences do not cause distortions in the individual assessments of mathematic competence, although variance in the outcome is higher due to interviewer presence.

Figure 1. Residuals of interviewers with corresponding posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).



Impact of Influential Interviewers

The impact of the identified interviewers were examined by evaluating (1) the number of missing values in the administered mathematic test, and (2) participation rates in a subsequent competence assessment about five to six years later. First, respondents tested by interviewers with significantly higher residual estimates (M = 2.18, SD = 2.23, N = 223) had significantly ($p < .05$) less missing values on the competence test as compared to respondents tested by non-outlying interviewers (M = 3.15, SD = 3.67, N = 4406); $t(286.72) = 6.15$, $p < .001$, $d = 0.27$. In contrast, for respondents tested by interviewers with significantly lower residual estimates (M = 3.42, SD = 3.68, N = 510), no significant difference in the number of missing values was found; $t(631.99) = -1.57$, $p = .118$, $d = 0.07$. Second, we compared the average participation rates for the subsequent competence assessment. For the respondent group tested by interviewers with non-outlying residual estimates 37.20 percent did not participate at the next assessment as compared to 47.98 percent for the interviewers with significantly higher residual estimates and 40.78 percent for the interviewers with significantly lower residual estimates. These differences in response rates were significant at $z(1) = 11.21$, $p < .001$, for the interviewers with significantly higher residual estimates, but not significant for interviewers with significantly lower residual estimates, $z(1) = 3.00$, $p = .083$.

Discussion

Interviewers play a decisive role in social surveys and educational large-scale assessments. Particularly, in household studies that visit respondents in their private homes interviewers have a great responsibility and need to create standardized settings while administering questionnaires and achievement tests under comparable conditions. If specific interviewer behavior affects the responses of participants, the validity of the measured constructs might be called into question. Therefore, survey managers need to evaluate the interview process and identify interviewers with an undue impact on respondent behavior.

The present study examined interviewer effects on mathematical achievement in a German large-scale assessment. Our Bayesian estimation of higher-order random effects in adult mathematic achievement identified a considerable number of interviewers that exhibited pronounced effects on the competence measurement, while area effects were negligible. These interviewer effects can yield important consequences. For one, statistical analyses that ignore the multilevel structure, especially the clustering of respondents in different interviewers, might result in underestimation of standard errors and, consequently, in the overestimation of statistical significance of found effects (Durrant et al. 2010; Finch and Bolin 2017). As an alternative to multilevel modeling a Huber/White correction to obtain robust standard errors in statistical analysis is appropriate (Huber 1967; White 1982) if estimates of second-level standard errors are biased. More information on that procedure can be found in Goldstein (2011) and Raudenbush and Bryk (2002).

Implications for Large-Scale Assessments

What are the implications of the presented results? First, practitioners engaged in large-scale assessments need to minimize interviewer effects on competence measurements. Even though large efforts are already invested into interviewer training and standardization of test situations, our results stress the need for further improvements, with the goal of achieving comparable settings for all test takers. Interviewer abilities are decisive in obtaining answers from different respondents that can be aggregated and compared across respondents to derive generalizable conclusions about population effects. Nevertheless, considering the sensitivity of individual assessment scores in the presented study, the presence of interviewer variance does not lead to bias in individual assessments.

To reduce interviewer variance on population effects, educational measurement could be improved by switching to an institutionalized setting that tests all respondents in highly standardized test centers. Further studies are needed that compare adult competence measurements in both individual and institutional settings. This might give invaluable insight

into interviewer effects introduced by different modes of administration. In comparison, large-scale educational data administered to students in a classroom setting found less than 1 percent of interviewer variance (Lüdtke et al. 2007).

Second, we presented a versatile methodological approach to empirically quantify interviewer effects on competence measurement. Bayesian analyses of cross-classified multilevel IRT models allowed us to disentangle interviewer from regional effects. Moreover, by investigating posterior draws of interviewer level random effect structures, inferences about effects from specific interviewers on competence testing can be made. Our study found that respondents interviewed by interviewers with significantly higher residual estimates had, in comparison to the respondents interviewed by non-outlying interviewers, significantly lower missing values in the competence test, but also lower participation rates at the subsequent measurement occasion. For respondents that were interviewed by interviewers with significantly lower residual estimates, no significant differences were found. So far, the precise reasons why outlying interviewers exerted these effects are unclear. It might be the case that they (unintentionally) interfered with the competence assessment (e.g., gave unrequested assistance) that bothered respondents and refrained them from further participation. Survey managers can use this approach as a tool for intervention, by having regular updates of the posterior distributions during data collection. As the posterior distributions point to interviewers with a significant effect on the survey measures, these interviewers can be additionally trained. Furthermore, to minimize the relatedness between interviewer- and area-clusters, we recommend a sampling design where each interviewer works in more than one area and each area is visited by more than one interviewer.

Limitations and Directions for Future Research

As a limitation of our study, a considerable number of interviewers worked in only one sample area and unobservable confounding of interviewer and area effects exists. Hence, the dependencies cannot be fully distinguished by the measurement of separated random

effect structures. Consequently, our results might be slightly distorted as compared to results obtained from a design, where interviewers are randomly distributed across areas. Nevertheless, this design limitation is common to national surveys. Moreover, a recent simulation study (Vassallo et al. 2017) found that three regions per interviewer are sufficient dispersion to obtain accurate estimates.

Interviewers were assigned to respondents based on spatial proximity of the living addresses, limiting the validity of our results. The multilevel cross-classified model assumes that the residual structures (σ^2_u, Σ_c, and Σ_g) are independent, but given the design of the study the interviewer-to-area distribution is not random. Hence, the assumption of independent residual structures (Σ_c, and Σ_g) is violated by the design of interviewer-to-respondent allocation. Even though we assume a limited interpenetrated design as being sufficient to disentangle interviewer and area clusters as sources of variance, unobservable confounding remains. In a fully interpenetrated design, where interviewers are assigned randomly to respondents, differences in interviewer means would allow a causal interpretation. With such a design, differences in interviewer means would reflect true differences in interviewer behavior. Unfortunately, a random allocation of interviewers across areas implies high costs for nationwide studies.

Although the presented results highlighted the influence of interviewer behavior on competence measurements, more research is needed to identify potential predictors of non-ignorable interviewer effects. In our analyses, the heterogeneity of survey estimates across interviewers was not related to observed interviewer characteristics. Therefore, future research should examine additional background information on the interviewers and the test administration process to understand the origin of interviewer effects. This might help alleviate respective effects by adapting the study design or improving the recruitment and training of interviewers. Moreover, our approach of identifying influential interviewers could be refined. Finding some interviewers with larger random effects might be expected because

of the assumption of multivariate normally distributed intercepts on the second level of our multilevel model (Finch and Bolin 2017). So far, it is unknown whether significantly outlying interviewers also adversely affect the validity of the competence estimates in large-scale educational assessments. Future research needs to develop measures that give further insights into the amount of deviation per interviewer; especially measures of severity for found outliers are needed.

Conclusion

The presented analyses reemphasize the conclusion of Schaeffer and colleagues (2010): interviewers are important in complex samples, helpful when critical response rates are expected, and especially useful in complex measurements. Therefore, we recommend intensified training and close monitoring for all tasks performed by the interviewers in the field, starting from respondent recruitment and persuasion for survey participation up to the standardized test administration. For this purpose, our Bayesian multilevel approach can be implemented to identify conspicuous interviewers during the ongoing data collection process.

References

Asparouhov, T., and Muthén, B. (2016), "General random effect latent variable modeling: Random subjects, items, contexts, and parameters," in *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*, eds. J. Harring, L. Stapleton and S. Beretvas, ch. 6, pp. 163-129. Charlotte, NC: Information Age Publishing.

Beullens, K. and Loosveldt, G. (2014), "Interviewer Effects on Latent Constructs in Survey Research," *Journal of Survey Statistics and Methodology*, 2, 433–458.

Beullens, K. and Loosveldt, G. (2016), "Interviewer Effects in the European Social Survey," *Survey Research Methods*, 10, 103-118.

Blossfeld, H.-P., Roßbach, H.-G. and von Maurice, J. (2011), "Education as a Lifelong Process - The German National Educational Panel Study (NEPS)," *Zeitschrift für Erziehungswissenschaft:* Special Issue 14.

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.

Brunton-Smith, I., Sturgis, P. and Leckie, G. (2016), "Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location-scale model," *Journal of the Royal Statistical Society. Series A, 180*, 551–568.

Collins, M. (1980), "Interviewer variability: a review of the problem," *Journal of the Market Research Society*, 22, 77-95.

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.

Fellenberg, F., Sibberns, H., Jesske, B. and Hess, D. (2016), "Quality assurance in the context of data collection", in *Methodological issues of longitudinal surveys: The example of*

the National Educational Panel Study, eds. H.-P. Blossfeld, J. von Maurice, M. Bayer and J. Skopek, vol 1, ch. 5, pp. 579–593. Wiesbaden: Springer.

Finch, W. H. and Bolin, J. E. (2017), *Multilevel Modeling using Mplus,* Boca Raton: Champan and Hall – CRC.

Fox, J.-P. and Glas, C. A. W. (2001), "Bayesian estimation of a multilevel IRT model using gibbs sampling," *Psychometrika*, 66, 271-288.

Fox, J.-P. (2003), "Stochastic EM for Estimating the Parameters of a Multilevel IRT Model," *British Journal of Mathematical and Statistical Psychology*, 56, 65-81.

Fox, J.-P. (2010), *Bayesian Item Response Modeling: Theory and Applications,* New York: Springer.

Gabler, S. and Lahiri, P. (2009), "On the definition and interpretation of interviewer variability for a complex sampling design," *Survey Methodology*, 35, 85-99.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004), *Bayesian data analysis* (2nd ed.), Boca Raton: Chapman & Hall.

Gelman, A. and Rubin, D.B. (1992), "Inference from iterative simulation using multiple Sequences," *Statistical Science*, 7, 457-472.

Goldstein, H. (2011), *Multilevel Statistical Models* (4th ed.), Chichester: Wiley.

Goldstein, H. and Browne, W. (2005), "Multilevel Factor Analysis Models for Continuous and Discrete Data," in *Contemporary Psychometrics*, eds. A. Maydeu-Olivares and J.J. McArdle, Chap 14, pp 453-475. New Jersey: Lawrence Erlbaum Assoc. Publishers.

Groves, R. M. (1989), *Survey Errors and Survey Costs,* New York: Wiley.

Hammon, A., Zinn, S., Aßmann, C., and Würbach, A. (2016), "Samples, Weights, and Nonresponse: the Adult Cohort of the National Educational Panel Study (Wave 2 to 6)," *NEPS Survey Paper No. 7*. Bamberg, Germany: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Hansen, M. H., Hurwitz, W. N. and Bershad, M. A. (1961), "Measurement errors in census and surveys," *Bulletin of the International Statistical Institute*, 38, 351-374.

Hanushek, E. and Woessmann, L. (2008), "The Role of Cognitive Skills in Economic Development," *Journal of Economic Literature*, 46, 607-668.

Hartig, J., Klieme, E. and Leutner, D. (2008), *Assessment of competencies in educational contexts*. Göttingen: Hogrefe Publishing.

Haunberger, S. (2010), "The effects of interviewer, respondent and area characteristics on cooperation in panel surveys: a multilevel approach," *Quality & Quantity*, 44, 957-969.

Heckman, J., Stixrud, J. and Urzua, S. (2006), "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24, 411-482.

Hox, J. J. (1994), "Hierarchical regression models for interviewer and respondent effects," *Sociological Methods & Research*, 22, 300–318.

Hox, J. J. (2010), *Multilevel analyses: techniques and applications* (2nd ed.), New Jersey: Lawrence Erlbaum Assoc. Publishers.

Hox, J. J. and de Leeuw, E. D. (1994), "A comparison of nonresponse in mail, telephone, and face-to-face surveys. Applying multilevel modelling to meta-analysis," *Quality & Quantity*, 28, 329-344.

Hox, J. J., de Leeuw, E. D. and Kreft, I. I. G. (1991), "The Effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model", in *Measurement Errors in Surveys*, eds. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudman, pp. 439-462. New York: Wiley.

Huber, P. J. (1976), "The behavior of maximum likelihood estimates under nonstandard conditions", in *Proceedings of the Fifth Berkeley Symposium on Mathematical*

Statistics and Probability, Volume 1: Statistics, 221-233, University of California Press, Berkeley.

Jiao, H., Kamata, A., & Xie, C. (2016), "A multilevel cross-classified testlet model for complex item and person clustering in item response modeling," in *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications*, eds. J. Harring, L. Stapleton, and S. Beretvas, ch. 5, pp. 139-162. Charlotte, NC: Information Age Publishing.

Jordan, A.-K. and Duchhardt, C. (2013), "NEPS Technical Report for Mathematics—Scaling results of Starting Cohort 6–Adults," *NEPS Working Paper No. 32*. Bamberg: University of Bamberg, National Educational Panel Study.

Kamata, A. and Vaughn, B. K. (2011), "Multilevel IRT modeling," in *Handbook of advanced multilevel analysis*, eds J. J. Hox and J. K. Roberts, ch. 3, pp. 41-57. New York: Routledge.

Kaplan, D. (2014), *Bayesian Statistics for the Social Sciences*, New York: Guilford Press.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Lee, E. S., Forthofe, R. N. and Lorimor, R. J. (1989), *Analyzing Complex Survey Data*, Newbury Park: Sage.

Lee, S.Y. and Song, X.Y. (2004), "Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes," *Multivariate Behavioral Research*, 39, 653-686.

Lüdtke, O., Robitzsch, A., Trautwein, U., Kreuter, F. and Ihme, J.-M. (2007), "Are there test administrator effects in large-scale educational assessments? Using cross-classified multilevel analysis to probe for effects on mathematic achievement and sample attrition," *Methodology*, 3, 149-159.

Luo, W., & Kwok, O. (2012), "The Consequences of Ignoring Individuals' Mobility in Multilevel Growth Models: A Monte Carlo Study," *Journal of Educational and Behavioral Statistics*, 37, 31–56.

Maas, C. J. and Hox, J. J. (2004), "Robustness issues in multilevel regression analysis," *Statistica Neerlandica*, 58, 127–137.

Mahalanobis, P. C. (1946), "Recent experiments in statistical sampling in the Indian Statistical Institute," *Journal of the Royal Statistical Society. Series A*, 109, 325-378.

Moss, P. A., Girard, B. J. and Haniford, L. C. (2006), "Validity in Educational Assessment," *Review of Research in Education*, 30, 109-162.

Muthén, B. and Asparouhov, T. (2012), "Bayesian SEM: A more flexible representation of substantive theory," *Psychological Methods*, 17, 313-335.

Muthén, B. and Asparouhov, T. (2016), "Multi-Dimensional, Multi-Level, and Multi-Timepoint Item Response Modeling," in *Handbook of Item Response Theory*, eds. W. J. van der Linden, vol 1, ch. 8, pp. 527-539. Boca Raton: CRC Press.

Muthén, L.K. and Muthén, B.O. (1998-2017), *Mplus User's Guide* (8th ed.), Los Angeles, CA: Muthén and Muthén.

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., and Ehmke, T. (2013), "Modeling and assessing mathematical competence over the lifespan," *Journal for Educational Research Online*, 5, 80.

OECD (2004) *Learning for Tomorrow's World – First Results from PISA 2003*, available at https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/34002216.pdf (last accessed June 18, 2020).

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.

O'Muircheartaigh, C. and Campanelli, P. (1999), "A multilevel exploration of the role of interviewers in survey-nonresponse," *Journal of the Royal Statistical Society. Series A*, 163, 437-446.

Pickery, J. and Loosveldt, G. (2004), "A simultaneous analysis of interviewer effects on various data quality indicators with identification of exceptional interviewers," *Journal of Official Statistics*, 20, 77–89.

Pohl, S. and Carstensen, C. H. (2013), "Scaling of competence tests in the National Educational Panel Study – many questions, some answers, and further challenges," *Journal for Educational Research Online*, 5, 189-216.

Rasbash, J., & Goldstein, H. (1994), "Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model," *Journal of Educational and Behavioral statistics*, 19, 337-350.

Rasch, G. (1980), *Probabilistic models for some intelligence and attainment tests*, Chicago, IL: University of Chicago Press.

Raudenbush, S. W. (1993), "A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research," *Journal of Educational Statistics*, 18, 321-349.

Raudenbush, S. W. & Bryk, A. S. (2002), *Hierarchical linear models* (2nd ed.), Sage, Thousand Oaks, CA.

Rosenthal, R. (1967), "Covert communication in the psychological experiment," *Psychological Bulletin*, 67, 356–367.

Rosenthal, R. (1994), "Interpersonal expectancy effects: A 30-year perspective," *Current Directions in Psychological Science*, 3, 176-179.

Rosenthal, R. (2002), "Covert communication in classrooms, clinics, courtrooms, and cubicles," *American Psychologist*, 57, 839-849.

Schaeffer, N. C., Dykema, J. and Maynard, D. W. (2010), "Interviewers and Interviewing," in *Handbook of Survey Research*, eds. P. V. Marsden and J. D. Wright, vol. 2, ch. 4, pp. 437-479. Bingley UK: Emerald.

Schnell, R. and Kreuter, F. (2002), „Separating Interviewer and Sampling-point Effects," in American Statistical Association *Proceedings of the Section on Survey Research Methods*, pp. 3132–3133.

Schnell, R. and Kreuter, F. (2005), "Separating interviewer and sampling point effects," *Journal of Official Statistics*, 21, 389–410.

Shelton, J. T., Elliott, E. M., Eaves, S. D., and Exner, A. L. (2009), "The distracting effects of a ringing cell phone: An investigation of the laboratory and the classroom setting," *Journal of Environmental Psychology*, 29, 513-521.

Skrondal, A. and Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Boca Raton, FL: Chapman & Hall/CRC.

Skrondal, A. and Rabe-Hesketh, S. (2009), "Prediction in multilevel generalized linear models," *Journal of the Royal Statistical Society. Series A*, 172, 659–687.

Tourangeau, R. and Yan, T. (2007), "Sensitive questions in surveys," *Psychological Bulletin*, 133, 859.

Turner, M., Sturgis, P., Martin, D. and Skinner, C. (2014), "Can interviewer personality, attitudes and experience explain the design effect in face-to-face surveys?," in *Improving Survey Methods: Lessons from Recent Research*, eds. U. Engel, B. Jann, P. Lynn, A. Scherpenzeel and P. Sturgis, ch. 7, pp. 72-85. Abingdon: Routledge.

Van den Noortgate, W., De Boeck, P. and Meulders, M. (2003), "Cross-classification multilevel logistic models in psychometrics," *Journal of Educational and Behavioral statistics*, 28, 369-386.

.

Vassallo, R., Durrant, G. B. and Smith, P.W. (2017), "Separating interviewer and area effects by using a cross-classified multilevel logistic model: simulation findings and implications for survey design," *Journal of the Royal Statistical Society. Series A*, 180, 531-550.

Vassallo, R., Durrant, G. B. and Smith, P. W., and Goldstein, H. (2015), "Interviewer effects on non-response propensity in longitudinal surveys: a multilevel modelling approach," *Journal of the Royal Statistical Society. Series A*, 178, 83-99

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2007), "A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests," *Educational and Psychological Measurement*, 67, 219-238.

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. and Olson, K. (2010), "How much of interviewer variance is really nonresponse error variance?," *Public Opinion Quarterly*, 74, 1004–1026.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica,* 50, 1-25.

Disentangling Interviewer and Area Effects in Large Scale Educational Assessments using Cross-Classified Multilevel Item Response Models

Online-only supplementary material

List of Tables

List of Figures

Table S1. Summary statistics of selected variables by hierarchical level

Variable	*M*/ %	*SD*	Min.	Max.	Information on Recoding	Name in original Dataset
Respondent Level (N = 5,139)						
Age	51.41	10.96	25	72	Grand-mean centred	tx29000
Gender (female)	0.51	-	0	1		t700001
Migration Background (yes)	0.17	-	0	1		t400500
Highest CASMIN					Recoded into 3 binary variables, reference category is ‘secondary education‘	tx28101
no degree or lower secondary education	0.19	-				
secondary education	0.33	-				
matriculation standard	0.19	-				
graduate degree	0.29	-				
Employment Status (unemployed)	0.20	-	0	1		tx29060
Cultural capital (Number of books)	4.11	1.33	1 (0 to 10 books)	6 (more than 500)	Grand-mean centred	t34005a
Political Area Size	4.17	1.78	1 (below 2000 inhabitants)	7 (more than 500k inhabitants)	Grand-mean centred	tx80103
Interviewer Level (N = 200)						
Gender (female)	0.40	-	0	1		tx80301
Age					Recoded into 2 binary variables, ref. categ. is ‚below 50 years‘	tx80302
below 50 years	0.22	-				
50 to 65 years	0.58	-				
older than 65 year	0.20	-				
Educational Attainment					Recoded into 2 binary variables, ref. categ. is ‚lower secondary degree‘	tx80303
lower secondary degree	0.14	-				
secondary education	0.31	-				
matriculation standard	0.55	-				
Work experience as interviewer					Recoded into 3 binary variables, ref. categ. is ‚up to 2 years‘	tx80304
up to 2 years	0.15	-				
2 to 3 years	0.31	-				
4 to 5 years	0.25	-				
more than 5 years	0.29	-				

Table S2. Summary of area to interviewer distribution

Number of test administrations per interviewer	Number of areas (regional clusters/ strata) per interviewer								
	1	2	3	4	5	6	7	8	Sum
1	3								3
2	1								1
3	2								2
4	5			1					6
5	7								7
6	1	1							2
7	5	3							8
8	7	1	1						9
9	7	2							9
10	1		1						2
11	3	1							4
12	5	1							6
13	6		1						7
14	3	1	2						6
15	4	2							6
16	3								3
17	2	2			1				5
18	1	3	1						5
19	3	1							4
20	2		2						4
21	2	3							5
22	2	2							4
23	2	3	2	2					9
24	2	8	1						11
25	1	1	1						3
26	2	2			1				5
27			2						2
28	1	1		2					4
29		2							2
30			2						2
31		2	1						3
33	1			1					2
34		2	1	1					4
35		1	1						2
36				2					2
37		1							1
38			1						1
40			1						1
42				1		1			2
43				1					1
44			1						1
45	1								1
46				1	2			1	4
48				2					2
49			1						1
50			1						1
52				1					1
53				1					1
55							1		1
56			1					1	2
57				1					1
58				1					1
60	1			1					2
63				1					1
67				1					1
68			1		1				2
70					1				1
71					1				1
73					1				1
76						1			1
77					1		1		2
78			1						1
80				1					1
84							1		1
91					1				1
95					1				1
123						1			1
Sum	86	46	27	22	11	3	3	2	200

Table S3. Summary of interviewer to area distribution

Number of interviewer per area (strata)

Number of Interviews per Area	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	17	26	28	Sum
1	1																		1
7	1																		1
10	2																		2
12		1																	1
13	1																		1
14	2	1																	3
15	3	1	2																6
17	2	2	1																5
18		1																	1
19		1																	1
20	2	1																	3
21	1																		1
22	1		1																2
24					1														1
25		2	1																3
26	2	1																	3
28			1																1
29		1	1	1															3
30					1														1
31						1													1
33		1																	1
34				1															1
35				1															1
36			1																1
37		1																	1
38			1																1
39			2	1															3
40			1																1
42		1																	1
43		1		1															2
44				1	1														2
45			1																1
46		1	1																2
49			1																1
52				1	1														2
53				1															1
60				1															1
61							1												1
67						1													1
69						1													1
70					1														1
73				1		1													2
74							1												1
75						1													1
83				1															1
84												1							1
86									1										1
87									1										1
95							1												1
97														1					1
98								1											1
99										1									1
110												1							1
113									1										1
119											1								1
122									1										1
126											1								1
155													1						1
162														1					1
174															1				1
195																1			1
199													1						1
239																		1	1
360																	1		1
Sum	18	17	15	11	5	5	3	1	4	1	2	2	2	2	1	1	1	1	92

Table S4. Summary of interviewer to area distribution (German federal states)

Number of test administrations per interviewer	Number of visited German federal states per interviewer						
	1	2	3	4	5	6	Sum
1	3						3
2	1						1
3	2						2
4	5	1					6
5	7						7
6	1	1					2
7	7	1					8
8	9						9
9	8	1					9
10	1	1					2
11	3	1					4
12	5	1					6
13	6	1					7
14	4	2					6
15	6						6
16	3						3
17	4			1			5
18	2	2	1				5
19	4						4
20	2	1	1				4
21	5						5
22	3	1					4
23	5	2	2				9
24	10	1					11
25	1	1	1				3
26	2	3					5
27	1	1					2
28	2		2				4
29	1	1					2
30	2						2
31	2		1				3
33	1	1					2
34	2	1	1				4
35	1		1				2
36		1	1				2
37	1						1
38	1						1
40	1						1
42			2				2
43		1					1
44		1					1
45	1						1
46	1	1	1			1	4
48		2					2
49	1						1
50		1					1
52		1					1
53	1						1
55				1			1
56			2				2
57		1					1
58		1					1
60	1	1					2
63	1						1
67	1						1
68		2					2
70			1				1
71	1						1
73				1			1
76					1		1
77			2				2
78	1						1
80		1					1
84				1			1
91		1					1
95		1					1
123				1			1
Sum	133	41	19	5	1	1	200

As an example, for the highlighted rows in Tables S2 to S4 it is demonstrated how to read the presented information.

S2: It occurred two times, that 42 interviews were conducted per interviewer, one of these two interviewers worked in four areas and the other interviewer worked in six areas.

S3: It occurred two times that 46 interviews were realized per area and that in each of these two areas, the interviews were conducted in one of these regions by 2 different interviewers and in the other region by 3 different interviewers.

S4: Two interviewers visited 3 German Federal States and each interviewed 42 respondents. Furthermore, more than one-third of the interviewers worked in more than one German federal state (*n*= 75, 37.5 percent). Each interviewer worked on average in 1.52 German federal states (min = 1, max = 6, *SD* = 0.87).

Sensitivity of variance components to prior choice

Variance components in hierarchical models can be sensitive to the choice of priors (Gustafson, Hossain and MacNab 2006). We conducted sensitivity analyses for priors of the estimated latent factor variances of the first model. Setting an inverse gamma prior of *IG* (.001, .001) for the interviewer latent factor variance estimate, or for all three estimates of latent factor variances, did not change the results substantially compared to model 1 (see results in last column of the subsequent table S5). Using an inverse gamma prior of *IG* (1,1) led to an increase in estimated latent factor variances, compared to the variances obtained using the Mplus default setting of *IG* (-1,0). Surprisingly, using the *IG* (1,1) specification for all three latent factor variances increased the area variance to nearly the same amount as interviewer variance. Hence, the random area variance might be sensitive to the choice of the prior to some degree, whereas the random interviewer variance is rather robust. However, an inverse gamma prior specification is not recommended for near-zero variance parameters in hierarchical models (Gelman 2006). As the area variance parameter is close to zero in the default prior setting, this high value might show misspecification by using the alternative prior *IG* (1,1).

References:

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.*, 1, 515-534. doi:10.1214/06-BA117A. (Available online: http://www.stat.columbia.edu/~gelman/research/published/taumain.pdf)

Gustafson, P., Hossain, S., and MacNab, Y. (2006). Conservative Prior Distributions for Variance Parameters in Hierarchical Models. *The Canadian Journal of Statistics / La Revue Canadienne De Statistique*, 34(3), 377-390.

Table S5. Sensitivity of latent factor variance estimates to choice of prior

Parameter	**Interviewer latent factor variance only**		**All latent factor variances**		**Latent factor variances of Model 1**
	IG (1,1)	*IG* (.001, .001)	*IG* (1,1)	*IG* (.001, .001)	*IG* (-1, 0)
Interviewer Variance	0.056	0.029	0.057	0.029	0.030*
Area Variance	0.003*	0.004*	0.054	0.003	0.004*
Within Variance (first Level)	0.465*	0.422*	0.508	0.418	0.425*

*Parameters were estimated with an inverse gamma prior specification of *IG*(-1, 0).

Table S6. Random item effects for interviewer clusters (estimated with Mplus, Version 8)

Parameter	M	SD	Item variance across interviewers	95% PPI
Discrimination				
Item 1	0.848	0.041	0.041	(0.005, 0.104)
Item 2	0.994	0.048	0.021	(0.002, 0.086)
Item 3	0.835	0.045	0.067	(0.030, 0.125)
Item 4	0.784	0.044	0.020	(0.002, 0.068)
Item 5	1.206	0.054	0.059	(0.007, 0.154)
Item 6	1.292	0.061	0.070	(0.007, 0.179)
Item 7	0.869	0.047	0.100	(0.049, 0.177)
Item 8	0.846	0.048	0.059	(0.012, 0.131)
Item 9	1.053	0.049	0.038	(0.004, 0.113)
Item 10	0.622	0.041	0.079	(0.041, 0.138)
Item 11	1.159	0.048	0.030	(0.002, 0.101)
Item 12	0.944	0.061	0.102	(0.026, 0.213)
Item 13	1.325	0.060	0.093	(0.021, 0.202)
Item 14	0.800	0.043	0.045	(0.006, 0.107)
Item 15	0.594	0.049	0.149	(0.077, 0.255)
Item 16	1.329	0.063	0.108	(0.023, 0.236)
Item 17	0.670	0.036	0.036	(0.005, 0.091)
Item 18	1.021	0.056	0.091	(0.022, 0.196)
Item 19	1.063	0.050	0.036	(0.003, 0.115)
Item 20	1.167	0.086	0.019	(0.002, 0.097)
Item 21	1.459	0.075	0.182	(0.083, 0.335)
Threshold				
Item 1	-0.258	0.029	0.011	(0.002, 0.031)
Item 2	-1.049	0.038	0.018	(0.002, 0.053)
Item 3	0.939	0.038	0.058	(0.027, 0.102)
Item 4	-1.213	0.034	0.005	(0.001, 0.021)
Item 5	-0.484	0.036	0.006	(0.001, 0.022)
Item 6	0.535	0.042	0.042	(0.011, 0.086)
Item 7	-0.126	0.032	0.029	(0.008, 0.060)
Item 8	-1.117	0.034	0.006	(0.001, 0.026)
Item 9	-0.080	0.036	0.033	(0.010, 0.065)
Item 10	0.361	0.030	0.033	(0.011, 0.066)
Item 11	-0.185	0.038	0.047	(0.021, 0.084)
Item 12	0.912	0.046	0.035	(0.004, 0.089)
Item 13	-0.003	0.037	0.012	(0.002, 0.037)
Item 14	-0.638	0.032	0.029	(0.009, 0.057)
Item 15	0.834	0.033	0.032	(0.007, 0.071)
Item 16	-0.032	0.039	0.023	(0.003, 0.057)
Item 17	-0.008	0.030	0.032	(0.011, 0.061)
Item 18	-0.632	0.036	0.023	(0.004, 0.054)
Item 19	-0.466	0.038	0.039	(0.010, 0.084)
Item 20	-2.033	0.089	0.146	(0.056, 0.283)
Item 21	0.797	0.047	0.021	(0.002, 0.063)
Latent Trait Variance				
Observations	0.635	-		
Interviewer	0.068	0.012		

Note. *M* = posterior mean. *SD* = posterior standard deviation. Item variance across interviewers = the item-specific random effect variance across interviewers. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Table S7. Random item effects for interviewer clusters (estimated with R-package SIRT)

Parameter	M	SD	Item variance across interviewers	95% PPI
Discrimination				
Item 1	0.860	0.036	0.091	(0.057, 0.137)
Item 2	1.010	0.045	0.085	(0.053, 0.133)
Item 3	0.832	0.039	0.098	(0.064, 0.144)
Item 4	0.791	0.042	0.079	(0.050, 0.121)
Item 5	1.223	0.047	0.112	(0.068, 0.176)
Item 6	1.309	0.052	0.117	(0.068, 0.188)
Item 7	0.871	0.037	0.109	(0.072, 0.162)
Item 8	0.853	0.043	0.097	(0.060, 0.149)
Item 9	1.074	0.045	0.099	(0.060, 0.160)
Item 10	0.626	0.035	0.098	(0.063, 0.144)
Item 11	1.184	0.043	0.094	(0.059, 0.148)
Item 12	0.946	0.052	0.135	(0.081, 0.215)
Item 13	1.333	0.049	0.127	(0.077, 0.204)
Item 14	0.811	0.037	0.085	(0.053, 0.126)
Item 15	0.594	0.036	0.159	(0.100, 0.244)
Item 16	1.342	0.052	0.151	(0.086, 0.244)
Item 17	0.677	0.033	0.079	(0.051, 0.118)
Item 18	1.028	0.046	0.126	(0.076, 0.199)
Item 19	1.083	0.047	0.088	(0.055, 0.135)
Item 20	1.099	0.076	0.100	(0.057, 0.162)
Item 21	1.452	0.064	0.189	(0.109, 0.297)
Threshold				
Item 1	-0.301	0.022	0.048	(0.033, 0.067)
Item 2	-1.123	0.030	0.063	(0.042, 0.091)
Item 3	0.904	0.028	0.083	(0.058, 0.117)
Item 4	-1.274	0.029	0.048	(0.032, 0.071)
Item 5	-0.554	0.026	0.047	(0.032, 0.067)
Item 6	0.468	0.028	0.075	(0.050, 0.109)
Item 7	-0.174	0.023	0.062	(0.042, 0.089)
Item 8	-1.186	0.030	0.051	(0.034, 0.072)
Item 9	-0.134	0.025	0.064	(0.045, 0.088)
Item 10	0.331	0.023	0.064	(0.044, 0.089)
Item 11	-0.250	0.024	0.070	(0.048, 0.099)
Item 12	0.870	0.039	0.081	(0.052, 0.122)
Item 13	-0.078	0.024	0.054	(0.037, 0.077)
Item 14	-0.693	0.026	0.059	(0.042, 0.082)
Item 15	0.813	0.028	0.069	(0.046, 0.099)
Item 16	-0.102	0.025	0.062	(0.042, 0.089)
Item 17	-0.040	0.022	0.061	(0.042, 0.085)
Item 18	-0.694	0.028	0.062	(0.041, 0.089)
Item 19	-0.529	0.027	0.072	(0.048, 0.104)
Item 20	-2.078	0.077	0.162	(0.091, 0.260)
Item 21	0.729	0.034	0.070	(0.045, 0.102)
Latent Trait Variance				
Observations	0.635	0.011		
Interviewer	0.062	0.021		

Note. *M* = posterior mean. *SD* = posterior standard deviation. Item variance across interviewers = the item-specific random effect variance across interviewers. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Table S8. Absolute differences between values of Table S6 and Table S7

Parameter	**M**	**SD**	**Item variance across interviewers**	**95% PPI**
Discrimination				
Item 1	0.012	0.005	0.050	(0.052, 0.033)
Item 2	0.016	0.003	0.064	(0.051, 0.047)
Item 3	0.003	0.006	0.031	(0.034, 0.019)
Item 4	0.007	0.002	0.059	(0.048, 0.053)
Item 5	0.017	0.007	0.053	(0.061, 0.022)
Item 6	0.017	0.009	0.047	(0.061, 0.009)
Item 7	0.002	0.010	0.009	(0.023, 0.015)
Item 8	0.007	0.005	0.038	(0.048, 0.018)
Item 9	0.021	0.004	0.061	(0.056, 0.047)
Item 10	0.004	0.006	0.019	(0.022, 0.006)
Item 11	0.025	0.005	0.064	(0.057, 0.047)
Item 12	0.002	0.009	0.033	(0.055, 0.002)
Item 13	0.008	0.011	0.034	(0.056, 0.002)
Item 14	0.011	0.006	0.040	(0.047, 0.019)
Item 15	0.000	0.013	0.010	(0.023, 0.011)
Item 16	0.013	0.011	0.043	(0.063, 0.008)
Item 17	0.007	0.003	0.043	(0.046, 0.027)
Item 18	0.007	0.010	0.035	(0.054, 0.003)
Item 19	0.020	0.003	0.052	(0.052, 0.020)
Item 20	0.068	0.010	0.081	(0.055, 0.065)
Item 21	0.007	0.011	0.007	(0.026, 0.038)
Threshold				
Item 1	0.043	0.007	0.037	(0.031, 0.036)
Item 2	0.074	0.008	0.045	(0.040, 0.038)
Item 3	0.035	0.010	0.025	(0.031, 0.015)
Item 4	0.061	0.005	0.043	(0.031, 0.050)
Item 5	0.070	0.010	0.041	(0.031, 0.045)
Item 6	0.067	0.014	0.033	(0.039, 0.023)
Item 7	0.048	0.009	0.033	(0.034, 0.029)
Item 8	0.069	0.004	0.045	(0.033, 0.046)
Item 9	0.054	0.011	0.031	(0.035, 0.023)
Item 10	0.030	0.007	0.031	(0.033, 0.023)
Item 11	0.065	0.014	0.023	(0.027, 0.015)
Item 12	0.042	0.007	0.046	(0.048, 0.033)
Item 13	0.075	0.013	0.042	(0.035, 0.040)
Item 14	0.055	0.006	0.030	(0.033, 0.025)
Item 15	0.021	0.005	0.037	(0.039, 0.028)
Item 16	0.070	0.014	0.039	(0.039, 0.032)
Item 17	0.032	0.008	0.029	(0.031, 0.024)
Item 18	0.062	0.008	0.039	(0.037, 0.035)
Item 19	0.063	0.011	0.033	(0.038, 0.020)
Item 20	0.045	0.012	0.016	(0.035, 0.023)
Item 21	0.068	0.013	0.049	(0.043, 0.039)
Latent Trait Variance				
Observations	-	-		
Interviewer	0.006	0.009		

Table S9. Random item effects for area clusters (estimated with Mplus, Version 8)

Parameter	M	SD	Item variance across areas	95% PPI
Discrimination				
Item 1	0.865	0.043	0.025	(0.003, 0.082)
Item 2	1.018	0.051	0.017	(0.002, 0.068)
Item 3	0.833	0.048	0.045	(0.011, 0.108)
Item 4	0.809	0.047	0.019	(0.002, 0.072)
Item 5	1.212	0.054	0.039	(0.004, 0.118)
Item 6	1.303	0.059	0.033	(0.003, 0.119)
Item 7	0.828	0.051	0.074	(0.029, 0.150)
Item 8	0.850	0.048	0.025	(0.002, 0.082)
Item 9	1.070	0.049	0.019	(0.002, 0.077)
Item 10	0.593	0.048	0.074	(0.030, 0.144)
Item 11	1.170	0.049	0.024	(0.002, 0.093)
Item 12	0.938	0.059	0.023	(0.002, 0.098)
Item 13	1.366	0.068	0.087	(0.020, 0.207)
Item 14	0.789	0.043	0.025	(0.003, 0.074)
Item 15	0.512	0.053	0.101	(0.045, 0.196)
Item 16	1.329	0.058	0.031	(0.003, 0.108)
Item 17	0.682	0.040	0.030	(0.004, 0.081)
Item 18	1.025	0.056	0.047	(0.006, 0.128)
Item 19	1.063	0.055	0.039	(0.003, 0.117)
Item 20	1.130	0.090	0.062	(0.005, 0.202)
Item 21	1.475	0.094	0.267	(0.128, 0.515)
Threshold				
Item 1	-0.254	0.031	0.005	(0.001, 0.020)
Item 2	-1.036	0.039	0.011	(0.001, 0.038)
Item 3	0.904	0.043	0.063	(0.027, 0.118)
Item 4	-1.199	0.035	0.007	(0.001, 0.030)
Item 5	-0.469	0.040	0.017	(0.003, 0.046)
Item 6	0.577	0.048	0.036	(0.006, 0.091)
Item 7	-0.082	0.032	0.011	(0.001, 0.034)
Item 8	-1.097	0.037	0.012	(0.002, 0.037)
Item 9	-0.094	0.039	0.016	(0.002, 0.046)
Item 10	0.387	0.032	0.022	(0.005, 0.054)
Item 11	-0.145	0.040	0.019	(0.004, 0.046)
Item 12	0.912	0.048	0.025	(0.003, 0.072)
Item 13	0.009	0.040	0.007	(0.001, 0.027)
Item 14	-0.627	0.034	0.016	(0.003, 0.043)
Item 15	0.847	0.034	0.016	(0.002, 0.048)
Item 16	-0.033	0.042	0.013	(0.002, 0.041)
Item 17	-0.002	0.030	0.009	(0.001, 0.030)
Item 18	-0.623	0.041	0.025	(0.006, 0.058)
Item 19	-0.451	0.038	0.014	(0.002, 0.047)
Item 20	-1.956	0.081	0.053	(0.008, 0.146)
Item 21	0.812	0.053	0.023	(0.002, 0.080)
Latent Trait Variance				
Observations	0.632	-		
Area	0.035	0.010		

Note. *M* = posterior mean. *SD* = posterior standard deviation. Item variance across areas = the item-specific random effect variance across areas. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Table S10. Random item effects for area clusters (estimated with R-package SIRT)

Parameter	M	SD	Item variance across areas	95% PPI
Discrimination				
Item 1	0.875	0.039	0.085	(0.052, 0.132)
Item 2	1.042	0.048	0.084	(0.052, 0.133)
Item 3	0.833	0.041	0.094	(0.057, 0.144)
Item 4	0.823	0.045	0.086	(0.052, 0.141)
Item 5	1.210	0.048	0.099	(0.059, 0.158)
Item 6	1.320	0.052	0.101	(0.059, 0.171)
Item 7	0.836	0.039	0.102	(0.064, 0.154)
Item 8	0.852	0.044	0.081	(0.051, 0.128)
Item 9	1.096	0.046	0.090	(0.054, 0.145)
Item 10	0.599	0.036	0.108	(0.067, 0.165)
Item 11	1.177	0.045	0.091	(0.056, 0.147)
Item 12	0.972	0.055	0.107	(0.062, 0.172)
Item 13	1.375	0.051	0.123	(0.072, 0.203)
Item 14	0.787	0.037	0.079	(0.049, 0.125)
Item 15	0.508	0.038	0.133	(0.080, 0.202)
Item 16	1.355	0.053	0.102	(0.061, 0.165)
Item 17	0.678	0.034	0.080	(0.051, 0.127)
Item 18	1.041	0.047	0.106	(0.063, 0.171)
Item 19	1.085	0.047	0.094	(0.058, 0.151)
Item 20	1.083	0.075	0.132	(0.072, 0.231)
Item 21	1.452	0.064	0.257	(0.147, 0.410)
Threshold				
Item 1	-0.293	0.023	0.049	(0.033, 0.070)
Item 2	-1.104	0.032	0.061	(0.040, 0.089)
Item 3	0.865	0.027	0.093	(0.061, 0.141)
Item 4	-1.250	0.032	0.059	(0.039, 0.087)
Item 5	-0.537	0.027	0.063	(0.041, 0.092)
Item 6	0.525	0.030	0.082	(0.052, 0.127)
Item 7	-0.115	0.024	0.055	(0.036, 0.080)
Item 8	-1.158	0.031	0.057	(0.038, 0.084)
Item 9	-0.171	0.026	0.062	(0.041, 0.093)
Item 10	0.359	0.024	0.062	(0.040, 0.089)
Item 11	-0.198	0.025	0.060	(0.040, 0.086)
Item 12	0.881	0.039	0.080	(0.050, 0.125)
Item 13	-0.069	0.025	0.054	(0.036, 0.079)
Item 14	-0.681	0.025	0.060	(0.040, 0.086)
Item 15	0.841	0.029	0.066	(0.043, 0.098)
Item 16	-0.118	0.026	0.060	(0.040, 0.089)
Item 17	-0.022	0.024	0.057	(0.037, 0.081)
Item 18	-0.691	0.028	0.065	(0.042, 0.097)
Item 19	-0.517	0.029	0.066	(0.043, 0.097)
Item 20	-2.034	0.068	0.109	(0.064, 0.178)
Item 21	0.734	0.035	0.087	(0.054, 0.133)
Latent Trait Variance				
Observations	0.632	0.011		
Area	0.031	0.023		

Note. *M* = posterior mean. *SD* = posterior standard deviation. Item variance across areas = the item-specific random effect variance across areas. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Table S11. Absolute differences between values of Table S9 and Table S10

Parameter	M	SD	Item variance across areas	95% PPI
Discrimination				
Item 1	0.010	0.004	0.060	(0.049, 0.050)
Item 2	0.024	0.003	0.067	(0.050, 0.065)
Item 3	0.000	0.007	0.049	(0.046, 0.036)
Item 4	0.014	0.002	0.067	(0.050, 0.069)
Item 5	0.002	0.006	0.060	(0.055, 0.040)
Item 6	0.017	0.007	0.068	(0.056, 0.052)
Item 7	0.008	0.012	0.028	(0.035, 0.004)
Item 8	0.002	0.004	0.056	(0.049, 0.046)
Item 9	0.026	0.003	0.071	(0.052, 0.068)
Item 10	0.006	0.012	0.034	(0.037, 0.021)
Item 11	0.007	0.004	0.067	(0.054, 0.054)
Item 12	0.034	0.004	0.084	(0.060, 0.074)
Item 13	0.009	0.017	0.036	(0.052, 0.004)
Item 14	0.002	0.006	0.054	(0.046, 0.051)
Item 15	0.004	0.015	0.032	(0.035, 0.006)
Item 16	0.026	0.005	0.071	(0.058, 0.057)
Item 17	0.004	0.006	0.050	(0.047, 0.046)
Item 18	0.016	0.009	0.059	(0.057, 0.043)
Item 19	0.022	0.008	0.055	(0.055, 0.034)
Item 20	0.047	0.015	0.070	(0.067, 0.029)
Item 21	0.023	0.030	0.010	(0.019, 0.105)
Threshold				
Item 1	0.039	0.008	0.044	(0.032, 0.050)
Item 2	0.068	0.007	0.050	(0.039, 0.051)
Item 3	0.039	0.016	0.030	(0.034, 0.023)
Item 4	0.051	0.003	0.052	(0.038, 0.057)
Item 5	0.068	0.013	0.046	(0.038, 0.046)
Item 6	0.052	0.018	0.046	(0.046, 0.036)
Item 7	0.033	0.008	0.044	(0.035, 0.046)
Item 8	0.061	0.006	0.045	(0.036, 0.047)
Item 9	0.077	0.013	0.046	(0.039, 0.047)
Item 10	0.028	0.008	0.040	(0.035, 0.035)
Item 11	0.053	0.015	0.041	(0.036, 0.040)
Item 12	0.031	0.009	0.055	(0.047, 0.053)
Item 13	0.078	0.015	0.047	(0.035, 0.052)
Item 14	0.054	0.009	0.044	(0.037, 0.043)
Item 15	0.006	0.005	0.050	(0.041, 0.050)
Item 16	0.085	0.016	0.047	(0.038, 0.048)
Item 17	0.020	0.006	0.048	(0.036, 0.051)
Item 18	0.068	0.013	0.040	(0.036, 0.039)
Item 19	0.066	0.009	0.052	(0.041, 0.050)
Item 20	0.078	0.013	0.056	(0.056, 0.032)
Item 21	0.078	0.018	0.064	(0.052, 0.053)
Latent Trait Variance				
Observations	-	-		
Area	0.004	0.013		

Interpretation of random item effects for interviewer and area clusters

Two separate hierarchical item response models with random discrimination and threshold effects were estimated. Furthermore, the results obtained from using the software Mplus (Version 8) were confirmed by using the function *mcmc.2pno.ml* from the R-Package sirt (Robitzsch 2019). We obtained comparable results from both programs for discrimination and threshold parameters as well as their item variances across interviewers or across areas (Table S8 and Table S11). The average deviation across interviewers for item variances between both programs was $M = 0.042$, $SD = 0.020$ ($Min = 0.007$, $Max = 0.081$) for the discrimination parameter and for the threshold parameter item variances the difference was $M = 0.036$, $SD = 0.008$ ($Min = 0.016$, $Max = 0.049$). Across areas, the average deviation for item variances between programs was $M= 0.055$, $SD = 0.018$ ($Min = 0.010$, $Max = 0.084$) for the discrimination parameter and for the threshold parameter item variances, the average difference was $M = 0.047$, $SD = 0.007$ ($Min = 0.030$, $Max = 0.064$).

The results for item variance across interviewer clusters (Table S6 and Table S7) and for item variance across area clusters (Table S9 and Table S10) are presented. Item discrimination and difficulty (threshold) parameters are depicted as well as the uncertainty which is given by the posterior standard deviation. Furthermore, random item effects at the interviewer level (Table S6 and Table S7) and area level (Table S9 and Table S10) with respective standard deviations are presented. In addition, 95% posterior probability intervals are given to evaluate significant deviations of the discrimination and threshold parameters. All estimated random effects at the interviewer level significantly deviate from zero when examining the 95% posterior probability interval. Nevertheless, it must be considered that variance estimates cannot become negative and in effect the probability interval will never include zero.

We assume that no strong violation of the measurement invariance is present. The share of variance in the latent trait across interviewers was 9.7 percent using Mplus (see last

two rows of Table S6) and 8.9 percent using the sirt-package for estimation (see last two rows of Table S7). The average variances of item parameters among interviewers across all items was 0.051 (average of item variances in Table S6; average of discrimination item variances was 0.069; average of threshold item variances was 0.032). The share of variance in the latent trait across areas was 5.2 percent using Mplus (see the last two row of Table S9) and 4.7 percent using the sirt-package for estimation (see the last two rows of Table S10). The average variance of item parameters was 0.036 (average of item variances in Table S9; average of discrimination item variances was 0.053; average of threshold item variances was 0.020). Hence, we assume that mathematic competence was measured as a unidimensional construct among interviewers and areas.

References:

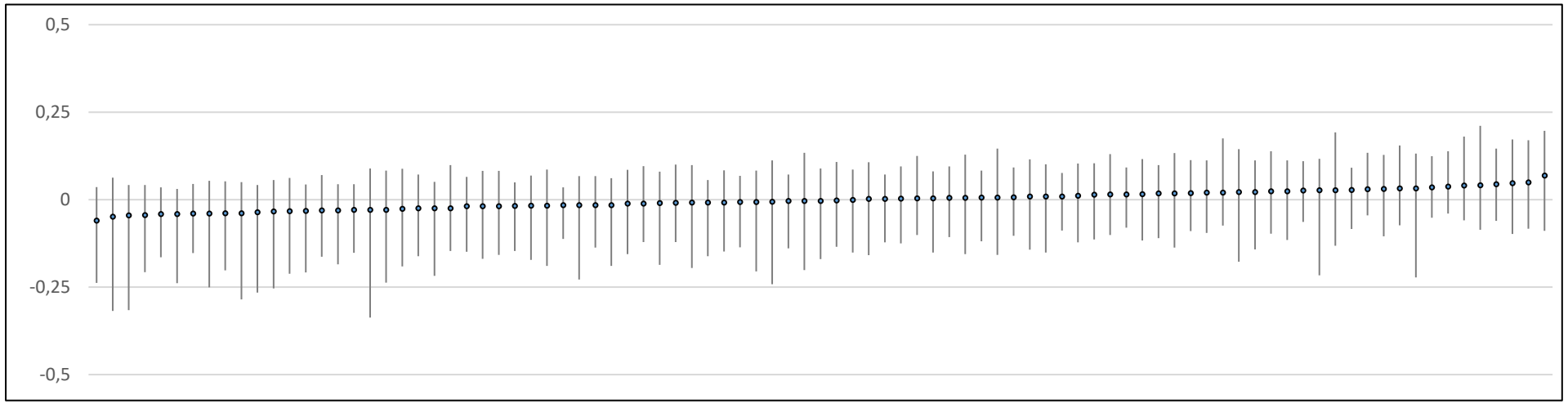
Robitzsch, A. (2019), *sirt: Supplementary Item Response Theory Models*. R package version 3.7-40, https://CRAN.R-project.org/package=sirt.

Table S12. Estimation results for the sample of interviewers having worked in at least two different regions (57 % of the interviewers)

	Model 1			**Model 2**			**Model 3**		
	M	*SD*	95% PPI	*M*	*SD*	95% PPI	*M*	*SD*	95% PPI
Fixed effects									
Age				-0.165	0.016	(-0.197, -0.134)	-0.165	0.016	(-0.195, -0.133)
Gender (ref. male)				-0.319	0.013	(-0.346, -0.293)	-0.318	0.014	(-0.345, -0.291)
Migration Background (ref. no)				-0.064	0.015	(-0.092, -0.035)	-0.064	0.014	(-0.093, -0.036)
Educational Attainment (ref. secondary education)									
no degree or lower sec. degree				-0.144	0.017	(-0.177, -0.111)	-0.144	0.017	(-0.178, -0.111)
matriculation standard				0.177	0.016	(0.145, 0.209)	0.177	0.016	(0.145, 0.208)
graduate degree				0.351	0.017	(0.318, 0.383)	0.351	0.017	(0.318, 0.383)
Employment status (ref. employed)				-0.056	0.015	(-0.087, -0.026)	-0.056	0.016	(-0.086, -0.026)
Cultural capital				0.160	0.017	(0.127, 0.193)	0.160	0.017	(0.127, 0.192)
Political Area Size				-0.043	0.021	(-0.084, -0.004)	-0.044	0.021	(-0.085, -0.003)
Interviewer Level Covariates									
Gender (ref. male)							-0.075	0.107	(-0.282, 0.141)
Age (ref. up to 49 years)									
50 to 65 years							-0.070	0.134	(-0.329, 0.194)
older than 65 years							0.103	0.136	(-0.166, 0.361)
Educational Attainment (ref. lower sec. degree)									
Secondary education							0.227	0.160	(-0.102, 0.521)
Matriculation standard							0.039	0.159	(-0.284, 0.338)
Work experience as interviewer (ref. up to two years)									
2 to 3 years							0.064	0.172	(-0.276, 0.394)
4 to 5 years							0.150	0.157	(-0.164, 0.454)
more than 5 years							0.000	0.170	(-0.330, 0.330)
Variance components of random effects									
Respondents	0.423	0.039	(0.355, 0.506)	0.243	0.024	(0.199, 0.290)	0.249	0.023	(0.208, 0.299)
Interviewers	0.029	0.007	(0.018, 0.045)	0.031	0.007	(0.021, 0.047)	0.031	0.007	(0.020, 0.048)
Areas	0.003	0.003	(0.000, 0.011)	0.001	0.001	(0.000, 0.006)	0.001	0.002	(0.000, 0.006)

Note. Standardized results are presented for fixed effects. *M* = posterior mean. *SD* = posterior standard deviation. PPI = posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).

Figure S1. Residuals of area clusters with corresponding posterior probability interval (2.5th and 97.5th percentile of the posterior distribution).



Design Effect for Interviewer and Area Clusters

There are two main consequences resulting from interviewer effects on survey outcomes: first, an increased variance of a statistic and second, a reduction in effective sample size. The impact of the first consequence on the measurement of mathematic achievement was tested by indicating interviewer and area variance proportions based on variance component testing (Intraclass Correlation Coefficient, ICC). It showed that most variance is attributable to interviewer clusters and a much smaller amount to sampling clusters, which is a finding shared by previous surveys (Brunton-Smith *et al.*, 2012; Brunton-Smith *et al.*, 2016; Durrant *et al.*, 2010; Schnell and Kreuter, 2005). The second consequence stems from the overall increase of variance due to the high interviewer effects, as both lead to a decrease in effective sample size. For this reason, even small effects per interviewer can have an undue impact on the data quality, especially if the caseload per interviewer is high (Collins, 1980; Hox, 1994; Kish, 1965; Schaeffer *et al.*, 2010).

The amount of dependence of resulting competence estimates on the test administrator can furthermore be expressed by the design effect. By this, the average size of interviewers' caseloads is considered additional to the ICC:

$$D_{eff} = 1 + (m - 1)\rho.$$

Thereby, m is the average number of test takers per interviewer and ρ is the ICC for all interviewers. Likewise, the design effect can be calculated for the area clusters, representing the effect of the two-stage sampling. Based on the intraclass correlation of Model 1 (see Table 1), the design effect for interviewer clusters amounts to 2.60 and to 1.44 for the area clusters. The design effect gives insight on how accurate the results are in comparison to a random sampling and at the same time it denotes how much larger the sample size must be to obtain the same precision in survey estimates (Schnell and Kreuter, 2005). Hence, it illustrates the increase in variance and also the decrease in effective sample size. For example, a design

effect of 2 reduces the effective sample size by half (Schaeffer *et al.*, 2010). As the design effect has no unit of measurement, its values are comparable across different survey estimates.

Sensitivity of interviewer variance (ICC) to outlying interviewers

As interviewers with deviating residuals introduce variance to the estimation of latent mathematic competence, we tested how much the intraclass correlation reduces when first, the most outlying interviewer with respective respondents and second, all outlying interviewers with respective respondents are excluded from the analysis. The estimation of our null model without the most outlying interviewer resulted in a reduced interviewer variance of 3.5 percent (in comparison to 6.6 percent of variance in Model 1 of Table 1), whereas the variance attributable to the respondents nesting in areas slightly increased to 1.1 percent (in comparison to 0.8 percent of variance in Model 1 of Table 1). Estimating the null model without all 12 outlying interviewers resulted in a further reduction of interviewer variance. The interviewer clusters now account for 0.9 percent of variance, with area clusters showing likewise a variance of 0.9 percent.

Sensitivity of interviewer residuals to group size

For most of the obtained interviewer residuals from our estimated multilevel IRT analyses, shrinkage to the general mean is expectable. If an interviewer interviewed a high number of respondents, posterior means resemble practically the intercept of separate regression estimations for this interviewer. Hence, the identification of exceptional interviewers might depend on the group size (the number of respondents per interviewer), also termed sensitivity of interviewer residuals to group size (Pickery and Loosveldt, 2004). We tested if the amount of residual deviation per interviewer cluster is correlated with the number of test administrations per interviewer. The correlation coefficient ($r = .074$, $p = .303$) does not indicate that the amount of uncertainty on the interviewer level depends on the size of the clusters.