Kroh, Martin; Winter, Florin; Schupp, Jürgen

Article

Using Person-Fit Measures to Assess the Impact of Panel Conditioning on Reliability

Public Opinion Quarterly

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Kroh, Martin; Winter, Florin; Schupp, Jürgen (2016) : Using Person-Fit Measures to Assess the Impact of Panel Conditioning on Reliability, Public Opinion Quarterly, ISSN 1537-5331, Oxford University Press, Oxford, Vol. 80, Iss. 4, pp. 914-942, http://dx.doi.org/10.1093/poq/nfw025

This Version is available at:
http://hdl.handle.net/10419/167599

Using Person-Fit Measures to Assess the Impact of Panel Conditioning on Reliability

Martin Kroh

(corresponding author)

Socio-Economic Panel, DIW Berlin

Mohrenstraße 58, 10117 Berlin, Germany, Phone +49-30-897678

and Humboldt-Universität zu Berlin, mkroh@diw.de

Florin Winter

Mohrenstraße 58, 10117 Berlin, Germany, Phone +49-30-897671

Mohrenstraße 58, 10117 Berlin, Germany, fwinter@diw.de

Jürgen Schupp

Socio-Economic Panel, DIW Berlin

Mohrenstraße 58, 10117 Berlin, Germany, Phone +49-30-897238

and Freie Universität Berlin, jschupp@diw.de

Running Header: The Impact of Panel Conditioning on Reliability

6,328 Words (excluding figures, tables, and references)

Author Affiliation

MARTIN KROH is Deputy Director of the Socio-Economic Panel Study (SOEP) at the German Institute for Economic Research (DIW Berlin) and a professor in the Department of Social Sciences at Humboldt-Universität zu Berlin, Berlin, Germany. FLORIN WINTER is a researcher at the Socio-Economic Panel Study (SOEP) at the German Institute for Economic Research (DIW Berlin), Berlin, Germany. JÜRGEN SCHUPP is Director of the Socio-Economic Panel Study (SOEP) at the German Institute for Economic Research (DIW Berlin) and a professor in the Department of Political and Social Sciences at Freie Universität Berlin, Berlin, Germany. The authors thank Deborah Bowen and Luisa Hilgert for their editorial support and the participants of the Panel Survey Methods Workshop 2014 in Ann Arbor for valuable comments. We also wish to thank the editors and three anonymous reviewers for their constructive and very helpful suggestions on earlier drafts. Since 2002, the Socio-Economic Panel Study has received continued funding as an infrastructure unit of the Leibniz Association through the Joint Science Conference by the Federal Government and the State of Berlin. Before that, SOEP was primarily funded through the German National Science Foundation (DFG).
*Address correspondence to: Martin Kroh, Socio-Economic Panel, DIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany; e-mail: mkroh@diw.de

Abstract

Panel conditioning has posed one of the main challenges to panel studies since their inception in the social sciences. Aside from the risk of reactivity to previous interviews, there is reason to expect that cumulative survey experience increases the reliability of data emanating from panel studies relative to cross-sectional surveys. This positive aspect of recurrent interviewing for data quality has been given relatively little attention in the empirical research to date. Drawing on observational data from 30 waves of the German Socio-Economic Panel (SOEP), we study the effect of individual survey experience on reliability, focusing on person-fit statistics from item-response models. The analysis documents that four years of survey experience produce a higher increase in person reliability than tertiary education compared to primary education.

Panel data enable researchers not only to study mobility and dynamic processes but also to test causal relationships more rigorously than is possible with simple cross-sectional data. However, since the introduction of panel studies in the social sciences, panel conditioning has presented one of the main limitations to repeated interviewing of the same individuals (e.g., Lazarsfeld and Fiske 1938; Lazarsfeld 1940).[1] Indeed, a large body of experimental research shows that past interview experience affects respondents' answers relative to a simultaneously interviewed control group of first-time respondents (Cantor 2008; Warren and Halpern-Manners 2012).[2] These effects of panel conditioning typically are portrayed as undermining the validity of the data in the subsequent interview (Holt 1989).

Several types of bias have been distinguished in the literature. For instance, the initial interview may induce respondents to gather information about the interview topics and to change their views and even their behavior before the next interview (Zaller and Feldman 1992; Sudman , Bradburn, and Schwarz 1996;Sikkel and Hoogendoorn 2008). Another form of bias may occur when respondents recall and repeat their answers from the initial interview in order to avoid contradictory answers (Waterton and Lievesley 1989 refer to this phenomenon as "freezing"). Similarly, experienced respondents may develop strategies to navigate efficiently through the questionnaire in order to reduce their survey burden (Bailar 1989). Thus, irrespective of whether respondents indeed change their attitudes or whether they change their response behavior, the first interview biases answers to the second interview. In the literature on experimental designs, these forms of panel conditioning are often referred to as problems of internal validity (Campbell and Stanley 1966).

[1] Panel attrition is usually considered the second obstacle to the use of panel data. While there exists a large body of literature studying non-response bias and methods of compensating for this by way of imputation, weighting, and model specification, surprisingly little research has been done on bias induced by panel conditioning and means of compensating for it.

[2] Interviewing respondents on a particular topic may not only affect their attitudes but also their subsequent behavior. For instance, Kraut and McConahay (1973) show that responding to an election study conducted prior to an election increases subsequent turnout. Similar effects of surveys on individual behavior are documented for consumer choice and health behavior (Warren and Halpern-Manners 2012).

Although repeated interviewing entails the risk of reactivity to the previous interviews, it may also have positive effects on the quality of the data provided by respondents. As early as 1938, Lazarsfeld and Fiske (p. 597) noted that "under certain circumstances, the statistical reliability of repeated interviews with panel members is greater than that of answers gained from a series of distinct samples." In the repeated interview situation, initial problems of question comprehension may have been clarified (Bailar 1989; Waterton and Lievesley 1989) and interviewees may be more versed in retrieving the relevant information (Feldman and Lynch 1988; Fazio 1989; Chandon, Morwitz, and Reinartz 2005). Moreover, after having gone through an initial interview, respondents may feel more familiar with the interviewer and the interview setting (Fowler 1995; Nancarrow and Cartwright 2007) and in some cases with fairly complex survey instruments (Basso,Lowery, Ghormley, and Bornstein 2001; see also Richardson and Robinson 1921).[3] The latter may reduce the erroneous use of response categories and possibly socially desirable responding. Overall, respondents in panel surveys may become more experienced and their answers therefore less error-prone with each successive survey wave. As a consequence, data quality could rise with each wave.

This positive impact on data quality with repeated interviewing has received relatively little attention in the empirical research. Exceptions are a study by Jagodzinski, Kühnel, and Schmidt (1987) and one study conducted by Sturgis, Allum, and Brunton-Smith (2009). Based on the notion that being interviewed induces respondents to reflect further on the respective issues after the interview, which as a consequence leads to more crystallized attitudes, they find that scale reliability for attitudinal multi-item constructs increases in subsequent waves.

[3] In fact, many types of complex psychological testing, such as measures of implicit association, respondents' first trial interviews are considered to be training routines that can be used to achieve acceptable levels of reliability.

The present paper draws on data from the Socio-Economic Panel (SOEP), an ongoing mixed-mode panel survey of now 30 annual waves, to study the effect of respondents' survey experience on person reliability. While previous research has often compared *scale* reliabilities across panel waves, w**e** estimate *person* reliability using person-fit-statistics from item-response models to facilitate a multivariate individual-level analysis of reliability (Meijer and Sijtsma 2001). This approach allows us to track changes in individual reliability of respondents who entered the ongoing survey at very different points in time over the panel waves—all while controlling for additional confounding factors such as aging, period effects, and changes in fieldwork procedures. Our empirical analysis documents that respondents provide more reliable data with every new wave of data collection.

1. Literature Review

In recent decades, scholars of survey research have developed sophisticated models of response behavior informed particularly by cognitive and social psychology (Schuman and Presser 1981; Zaller and Feldman 1992; Tourangeau, Rips, and Rasinski 2000; Schwarz 2007). Measurement error may occur at all stages of this multipart response process, starting with the comprehension of the question, the retrieval of relevant information from memory, the judgment of the latent answer, and finally the selection of a response category. The following stylized literature review lends support to the hypothesis that the survey experience of respondents is generally associated with lower levels of measurement error at the different phases of the response process.

The first stage, the *comprehension* of the survey task and question, may induce measurement error if the instructions, question, or response options are highly complex, vague, or unfamiliar to the respondent (Schuman and Presser 1981; Fowler 1992). This ambiguity is mitigated, however, when respondents are confronted with the same survey question

repeatedly (see Waterton and Lievesley 1989; Binswanger, Schunk, and Toepoel 2013), reflecting an increase of knowledge on topics they have been asked about previously (Toepoel, Das, and van Soest 2009).

In the *retrieval* stage of the response process, interviewees draw relevant information from their long- and short-term memory based on their understanding of the survey task. Survey experience presumably eases the retrieval process for respondents, as repeated interviewing increases the accessibility of relevant information. In line with this view, Bailar (1989) indicates that the number of errors in recalling retrospective events decreases across survey waves (see Traugott and Katosh 1979 for a related interpretation of their findings on voting behavior, and Porst and Zeifang 1987 and Jagodzinski, Kühnel, and Schmidt 1987 for the consistency of attitudes in panel surveys).

In the next stage of the response process, respondents integrate the retrieved material and form a *judgment*, i.e., a tentative answer on the basis of this combined set of information (Kahnemann and Tversky 1979). In an influential study, Krosnick and Alwin (1987) discuss the phenomenon of satisficing, that is, a strategy to minimize the cognitive costs associated with this decision process. The prevalence of providing just sufficiently accurate information to minimize individual survey burden (Krosnick, Narayan, and Smith 1996; Krosnick 1999) is particularly high if the difficulty of the survey task is disproportionately higher than the ability and motivation of respondents (Bradburn 1978, Holbrook, Green, and Krosnick 2003). As respondent ability may also be gained through familiarity with a topic and survey experience, one would expect less satisficing and less erroneous judgment with increasing levels of survey experience.

Finally, respondents *report* their latent judgment using the offered response categories. Measurement error at this stage may occur because respondents experience difficulties in translating a latent judgment into, for instance, a single category on a numerical response

scale. Again, these difficulties are less likely in persons with survey experience, as they become more familiar in dealing even with complex survey instruments (Basso et al. 2001). In sum, several studies suggest that survey experience has the potential to increase the quality of survey data at different stages of the response process: Survey experience diminishes ambiguities in question comprehension, facilitates the retrieval of relevant information, reduces the need of heuristics of judgment, and makes haphazard misreporting less likely. Thus survey experience improves respondent ability and has similar properties to the effort respondents invest in answering survey questions as well as to their cognitive ability in general and their familiarity with the researched topic in particular.

Data and Analyses

Testing the hypothesis of this paper that individuals' survey experience reduces measurement error in their survey responses requires a person-specific statistic of measurement error on the one hand and variation in person-specific survey experience on the other. Due to the absence of *ex ante* information on the true score for respondents' answers in most instances, we use statistics of the internal consistency of multi-item constructs, referred to as person-fit statistics, to evaluate the degree of measurement error in individual response patterns. We apply this method to the Socio-Economic Panel (SOEP) described below. This ongoing annual mixed-mode panel survey with currently 30 panel waves and regular refreshment samples allows us to study both cross-sectional and longitudinal differences in person reliability across levels of survey experience.

The Socio-Economic Panel (SOEP)

The SOEP is an ongoing annual household panel survey of about 30,000 adults and 10,000 children in 16,000 private households in Germany every wave (Wagner, Frick, and Schupp 2007). The target population includes all residents of Germany, irrespective of nationality.

Questionnaires are translated into five languages and available online. The SOEP started in 1984 and consists today of several area-based as well as register-based probability samples drawn between 1984 and 2013. The 15 subsamples of SOEP include amongst others three boost of migrants as well as five cross-sectional refreshment samples. Since 1998, SOEP has been gradually replacing PAPI (personal paper and pencil interviewing) with CAPI (computer-assisted personal interviewing) as the predominant mode of data collection. On demand, experienced panel households may also use SAQ (self-administered questionnaire). Wave 1 response rates range between 33 and 70 percent across SOEP samples (AAPOR RR1) and longitudinal response rates between two consecutive waves has hovered around 90 percent from wave 3 on (Kroh 2013). The SOEP questionnaire covers topics such as work, income, family relations, well-being, health, public opinions, and civic participation. In recent years, measures of personality traits have also become part of the SOEP questionnaire.

To estimate the fit of observed response patterns to some underlying response models, we use all the available, established multi-item constructs that have been surveyed by SOEP There are a total of 17 multi-item constructs surveyed repeatedly between 1992 and 2013 (see Table 1). The scale manual of SOEP lists all these instruments, full question wordings as well as their theoretical background and lists of references (see Richter,Metzing, Weinhardt, and Schupp 2013). Five of these measures are part of an inventory of personality traits), three relate to concepts such as trust and reciprocity, and three deal with individual health and well-being. Furthermore, we consider multi-item measures of risk aversion, two versions of external locus of control, tendency to forgive), anomie, and finally an instrument measuring occupational and job stress, more specifically the effort reward (im)balance in working conditions. Table 1 reports the number and labeling of items per construct and also the number of (ordered) response options ranging from 2 to 11 (no/yes and strongly dis/agree). Table 2 reports the number of replications for each multi-item construct in the past 20 years.

[INSERT TABLE 1 ABOUT HERE]

More than 65,000 adults were surveyed by SOEP between 1984 and 2013. This paper only uses annual waves in which SOEP surveys the aforementioned 17 multi-item constructs. This reduces the sample size to 49,522 persons surveyed in the years 1992 through 1997 and 2002 through 2013. In both the overall and in this reduced sample, the median number of interviews per adult respondent is about 10 observations. In each of the analyzed waves, we observe between one and eight multi-item constructs. With one exception, the analyzed instruments have been surveyed between two and seven times over the years (see Table 2). This results in total in a sample of more than 1 million estimates of person reliability (persons x waves x instruments) emanating from more than 324,000 interviews (persons x waves) in more than 49,000 respondents (persons).

[INSERT TABLE 2 ABOUT HERE]

Person-Fit Statistics

Survey researchers usually describe reliability as a function of respondents, interviewers, and instruments. While it is customary to report the scale reliability of instruments, it is less common to study the person reliability of individual respondents. Person-fit statistics refer to measures that can be used effectively to evaluate the consistency of individual response patterns (van Vaerenbergh and Thomas 2013). The underlying objective of person-fit statistics is to identify aberrant response patterns in multi-item measures—that is, response patterns that deviate from the estimated item response model. The parametric person-fit statistics used throughout this paper are used to establish the difference between the observed and the expected item scores over a number of items (Meijer and Sijtsma 2001, Karabatsos 2003 for binary response options; van Krimpen-Stoop and Meijer 2002, Dagohoy 2005, Conijn 2013 for polytomous items).

In the case of ordinal rating items that are used predominantly in the context of the present study, the probability of a person selecting a certain item score is determined by the person's latent value on the concept of interest, θ, the step-difficulties of item i at the k-th cut-point, δ_{ik}, and the discrimination parameter of item i, λ_i. θ thus captures the position of the individual respondent on the latent trait (e.g., "tendency to forgive"). δ_{ik} displays the position on the latent dimension for which the probability of choosing score j (e.g., "[1] Does not apply to me at all") and j+1 (e.g., "[2] Does not apply to me most of the time") for item i (e.g., "I tend to bear grudges") intersect. λ_i, the discrimination parameter, indicates the loading (sometimes called relevance) of item i for the measurement model of the latent dimension. The maximum number of response categories of item i is denoted as m_i. The probabilities of each item score, P_{ij}, can thus be described as follows:

$$P_{ij} = \frac{exp \sum_{k=0}^{j}(\lambda_i(\theta - \delta_{ik}))}{\sum_{j=0}^{m_i} exp \sum_{k=0}^{j}(\lambda_i(\theta - \delta_{ik}))}, j = 0,1, \dots, m_i$$

The estimated model with two parameters (difficulty and discrimination) and item-specific thresholds for each cut-point, which is sometimes referred to as partial credit model (Masters 1982), is to our knowledge the standard item-response model for ordinal rating items employing as few identification restrictions as possible. Table 1 reports difficulty and discrimination parameters for all 17 multi-item measures analyzed in this paper. The following paragraphs briefly describe person-fit statistics formally and also provide an illustration.

Based on these estimated parameters, we obtain the predicted probabilities for a factual response pattern and are able to compare them with the observed response pattern. For this purpose, we use the "l-person-fit statistic" (Levine and Rubin 1979 for dichotomous variables

and Drasgow, Levine, and Williams 1985 for polytomous items). In the subsequent formal description, we follow van Krimpen-Stoop and Meijer (2002).

In this model, the log-likelihood of a factual response pattern is represented by

$$l = \sum_{i=1}^{N} \sum_{j=0}^{m_i} d_j(x_i) \ln P_{ij}$$

with $d_j(x_i) = 1$ for factually observed item scores x_i and $d_j(x_i) = 0$ if otherwise. The likelihood of a person's expected response pattern is represented by

$$E(l) = \sum_{i=1}^{N} \sum_{j=0}^{m_i} P_{ij} \ln P_{ij}$$

Since some response patterns are associated with higher probability-differences than others, we use the standardized version of the "l-person-fit statistic"

$$l_z = \frac{l - E(l)}{[Var\ (l)]^{1/2}}$$

Figure 1 illustrates the person-fit of two different response patterns for the instrument "tendency to forgive" (Brown 2003). This instrument consists of four seven-point rating items. Item 2 and 3 are reversed; that is, high item scores indicate low tendency to forgive. The figures in between the response options report the estimates of item-step difficulty, δ_{ik}, i.e., the position of the response option on the latent scale "tendency to forgive." Item discrimination parameters, λ_i, or the relevance of each item for the overall construct, are given in parentheses.[4]

We select two exemplary response patterns (X=person 1 and O=person 2). The estimated level of forgivingness in both individuals is about one standard deviation below the mean in the sample (θ_X =-0.70 and θ_O=-0.72). While both response patterns are associated with the

[4] Discrimination parameters for item 1 in each multi-item instrument are restricted to λ=1.

nearly the same positon on the latent trait, the first one [X] fits the estimated response model well and the second [O] poorly (l_{zX}=1.4 and l_{zO}=-3.4).

[INSERT FIGURE 1 ABOUT HERE]

In the fitting-response pattern, the first interviewee consistently uses those categories that match her position on the latent trait given the model-specific parameters. Hence the item step difficulties δ_{ik} of the cut-points above and below the marked boxes (that use the same scaling as the latent trait) typically include the value of θ_X =-0.70. In these instances, the answers match the highest predicted probabilities of responses for a given θ-value. In the second response pattern [O], characterized as aberrant, the interviewee selects response options that do not always match her value on the latent trait of θ_O =-0.72. For instance, answers indicating that she tends to bear grudges (applies to me perfectly) and that she tries to forgive and forget when other people wrong her (applies to me perfectly) inconsistently suggest low levels of forgivingness in item 3 (below item step difficulty δ_{37} = -4.0) and high levels of forgivingness in item 4 (above item step difficulty δ_{47} = +3.3. In this example, the respondent conceivably confounded positively and negatively labeled items, which leads to the poorly fitting response pattern of person 2 [O].

Figure 2 plots the distribution of all 1,079,068 estimated person-fit scores, limiting the range between -2 and +2. High values indicate that the observed response pattern fit the predicted probabilities of the item response model, and low values indicate a bad fit. The left-skewed distribution suggests that person-fit statistics capture not only (normally distributed) random error but also systematically lower levels of fit in a larger segment of the sample. Table 2 reports the mean and standard deviation of person-fit estimates for each multi-item construct individually.

[INSERT FIGURE 2 ABOUT HERE]

1.1 Person Reliability

Interpreting these person-fit statistics as an indicator of person reliability—i.e., an ability or effort of respondents to provide accurate survey data—requires a certain conceptual stability in person-fit statistics across instruments and time. If there is no correlation between person-fit statistics, one would interpret the statistics as a function of truly random measurement error caused, for instance, by temporal lapse in concentration or haphazardly occurring errors. Indeed, De Leeuw and Hox (1994) report a rather weak correlation in their cross-sectional analysis of person-fit statistics obtained from different multi-item measures (see also Schmitt et al. 1999), suggesting that a bad or good fit is to a considerable extent a function of idiosyncratic influences.

In the context of the SOEP, person-fit statistics for different scales correlate moderately in a cross-sectional perspective typically between .05 and .20 (for comparable figures, see Conijn 2013; for higher correlations, see Woods, Oltmann, and Turkheimer 2008). Nonetheless, even this low inter-instrument correlation is sufficient to establish a Cronbach's Alpha value of person-fit statistics across the 17 instruments of .67. Temporal stability of person-fit for single instruments ranges between .44 and .54 (intraclass-correlation attributable to persons). In other words, person-fit for single instruments is relatively stable over time, and person-fit for different instruments at one point in time is moderately correlated (see also Conijn 2013). These figures are comparable to previous studies, and in line with this research, we consider person-fit statistics sufficiently consistent to interpret them as person reliability.

1.2 Model Choice

Table 3 lists altogether six models that differ on the one hand in estimation strategy and on the other hand in terms of the set of control variables. As to the model specification, we compare pooled OLS models with person robust standard errors (1,4), (person-specific) fixed effects

models (2,5), and mixed effects models distinguishing between person and interviewer random effects (3,6). While pooled OLS estimates the effect of experiences drawing on both between-respondent variation as well as within-respondent variation in experience and person reliability, the fixed effect model uses only within-person changes in experience and reliability to establish an effect. Thus fixed effects models, which control for individual differences in person reliability that already existed before the beginning of the panel, are usually considered the more rigorous research design for causal inference.

As to the set of control variables, we distinguish between models that only include wave and instrument effects (for reasons of clarity, the effect coefficients are not reported in Table 3) as well as indicators of the cumulative experience of respondents (1-3), and models that further contain control variables that may affect person reliability and growing survey experience at the same time (4-6).

The first set of models reports that the 1,079,068 observations of person reliability come from 49,522 respondents interviewed between 1992 and 2013. The fixed effects estimates suggest that roughly 23 percent of the variance in person reliability is a time-invariant property of individual respondents and the remaining 77 percent of person reliability is either subject to temporal changes in individuals or specific to the person-instrument constellations and to idiosyncratic errors (for comparable figures, see Conijn 2013).

Restricting the analyzed sample in the mixed effects models to the main interviewer of a respondent in the period between 1992 and 2013 thereby ignoring changes in interviewer allocation suggests that only 2 percent of the variance in person reliability is due the unit effect of the 1,296 interviewers and 8 percent due to the unit effect of respondents.[5]

[5] Due to occasional changes in interviewer allocation, a single interviewer may not only interview several respondents but also a single respondent may be interviewed by several interviewers over time. The relatively high stability of interviewer allocation in the SOEP, however, leads to non-convergence of cross-classified models in the present data. We therefore restrict the sample in Models 3 and 6 to the main interviewer of

[INSERT TABLE 3 ABOUT HERE]

To identify likely causal processes, we use three alternative indicators of experience: first, the cumulative number of interviews with a respondent in the SOEP; second, the cumulative number of interviews with a respondent by the same interviewer; and third, the cumulative number of interviews with a respondent using the same multi-item instrument. While the first indicator measures general survey experience, the second indicator captures the familiarity with an interviewer, and the third the familiarity with a specific multi-item instrument. The first two experience indicators range in principle between 1 and 30 waves of experience and are capped for simplicity at 5 to 9 and at 10 or more interviews. The last one ranges between 1 and 7 replications of a multi-item instrument and are capped at 4 and more interviews for a single instrument. Table 4 reports descriptive statistics of all regressors.

The most important finding of Table 3 is that general experience and experience with a single multi-item instrument increase person reliability robustly in all six model specifications, but familiarity with the interviewer only in the first pooled OLS model. Conversely, more rigorous models such as fixed effects and mixed person and interviewer effect models suggest that the effect of interviewer familiarity is spurious. The positive effect of growing general experience monotonically increases with every wave, although with declining marginal growth rates. The largest absolute gains in person reliability are achieved in the first four years of survey participation.[6]

[INSERT TABLE 4 ABOUT HERE]

Robustness Checks

respondents and estimate ordinary two-level hierarchical models of respondents being clustered in interviewers.

[6] Collapsing years of experience into broader categories does not change the substantive interpretation. Estimating effects for each year of experience separately produce the same pattern of decreasing marginal effects of survey experience on person reliability as reported in this table.

Based on our literature review of the correlates of person reliability, we add a number of control variables to the regression models to further establish the robustness of the estimated positive effects of survey experience (Models 4 – 6). Previous research suggests that five factors are particularly relevant for the explanation of person reliability: (a) cognitive ability; (b) survey effort; (c) response behavior; (d) traitedness; (e) demographics; and (f) fieldwork characteristics.

Note that the fixed effects model (Model 5), which uses within-respondent variation to estimate the effect of experience on person reliability by definition omits time-invariant covariates such as gender.[7] Note also that not all of these variables are measured in all waves of the panel. Table 4 reports the number of observations indicating that particular information regarding respondents' individual characteristics, such as cognitive test scores and levels of conscientiousness, is missing. We use multiple imputation by chained equations to replace all missing information. We use both cross-sectional as well as longitudinal information for the imputation and employ all of the SOEP data since 1984 for the imputation procedure. The imputation was performed on the dataset in the long structure (person x waves). To capture the temporal stability of the data, we consider lagged variables in the imputation. All estimates reported in Table 3 build on the imputed data. However, we restrict the sample of analysis to cases with valid information on the outcome variable of person reliability. Due to the large number of observations, we confine the analysis to 10 fully imputed data files. Alternative approaches dealing with missing data do not affect the substantive conclusion of the article.[8]

Cognitive Ability

[7] Educational attainment is almost time-invariant in many adults and due to the limited number of replications over time, we also treat personality traits and cognitive ability test as time-invariant properties of individuals.

[8] We tested, for instance, the listwise deletion of missing information with/out omitting regressors most affected by missing data. Also, we parameterized a missing-category for all categorical regressors (missing value dummy variables procedure).

Conijn (2013) finds that education of respondents positively correlates with person reliability and Knäuper et al (1997) show a positive association between tests of cognitive ability and the prevalence of "don't know" answers in surveys. We therefore simultaneously control for the effect of ability in form of education as well as two cognitive tests (word fluency test and symbol digit test; see Lang et al. 2007) when estimating the effect of survey experience. The results of Table 3 show that—as expected—indicators of (time-invariant) cognitive ability positively affect person reliability.

Survey Effort

Schmitt et al. (1999) show that test-taking motivation and conscientiousness correlate positively with person-fit. Similarly, McFarland and Ryan (2000) report that conscientiousness is negatively related to faking answers. Hence, motivational differences over the course of a longitudinal survey might generate a spurious correlation between survey experience and person reliability. This may be particularly true for response styles, such as straightlining, that may emerge through survey experience and that create a false impression of consistency in answers. Although this issue is mitigated by the fact that many of the multi-item constructs employed in the paper use reversed items, straightlining using middle categories of rating scales would still falsely suggest high person reliability.

We code straightlining as the absence of variability in three sets of items surveyed (bi-) annually by SOEP since 1984, namely, happiness in various domains (e.g., income, health), opinions on several issues (e.g., crime, job security) and finally, time use in different areas (e.g., sports activities, church). Based on 20 to 30 rating scales in each wave, we code straightlining as minimal variability of answers in these item batteries.[9] Moreover,

[9] To avoid confoundedness with our measure of person reliability, we use alternative sets of items (happiness, worries, time use) instead of the 17 analyzed multi-item constructs to code straightlining. Response times are unavailable for single items and we therefore cannot measure straightlining by response speed.

conscientiousness and the prevalence of item non-response serve as additional indicators of the survey effort of respondents.

Regression models reported in Table 3 show increasing levels of person reliability by straightlining and decreasing person reliability by the prevalence of item refusals. However, contrary to our expectation and in contrast to previous studies, we find lower levels of person reliability in very conscientious respondents.

Response Behavior

The ability and motivation of respondents may also spuriously suggest a positive effect of experience on person reliability by way of non-random panel attrition. To the extent that more motivated respondents participate longer in the panel survey, the longitudinal sample of respondents will gradually consist of more motivated interviewees, and hence person reliability will due to this self-selection grow wave by wave.

Our strategy to deal with nonrandom panel attrition is threefold: First, we use fixed-effect models to capture within-person change in person reliability that eliminate the generic factors of individuals associated with person reliability (Models 2 and 5). Second, the Online Appendix A to this paper reports the regression models of Table 3 for a balanced sample of respondents in their first ten years of survey participation. This reduced subsample shows highly similar estimates of survey experience on person-fit compared to the larger, unbalanced sample of all observations that may change in its composition over time. Third, Models 4 – 6 include a measure of the *prospective* number of panel waves in respondents to obtaining an effect of survey experience conditional on future panel attrition. This measure captures whether person reliability for given levels of experience is lower in respondents who will refuse to participate in the following wave of the survey compared to those who will participate in the survey for several more years.

Another aspect of response behavior that may artificially introduce an association between experience and person reliability is the reported extremity of latent traits. In composite measures of particularly few items, the extreme ends of the latent trait are only associated with a single response pattern while at the center of the latent trait several response patterns may produce the same latent value θ of the concept of interest. That is, respondents scoring the highest (and lowest) possible value on the latent trait will always have a very good person-fit, whereas respondents with medium value on the latent trait can either have low or high person-fit. To capture this effect, we add a categorical control variable to the analysis which distinguishes between mean values on the latent trait, moderately positive/negative, and extreme values.[10]

Results reported in Table 3 indicate that indeed the person reliability in wave t of respondents who consistently participate in future waves is higher compared to respondents who will drop out at t+1. Moreover, individual values at the extreme ends of the latent trait θ are generally associated with higher levels of person reliability.

Traitedness

Aberrant response patterns have also been identified in unstable personalities: McFarland and Ryan (2000) show that neuroticism was negatively related to person-fit statistics, and Woods, Oltmann, and Turkheimer (2008) report that persons with signs of pathological behavior show more aberrant response styles. Tellegen (1988) more generally uses the term "traitedness" to denote the consistency between individuals' personalities and their behavior. According to Reise and Waller (1993), this traitedness is also reflected in the degree of aberrant response patterns. We use respondents' level of emotional stability, the experienced prevalence of

[10] More detailed regression analysis of the association between values on latent traits and person-fit statistics in principle confirm that for most multi-item instruments used in the analysis of this paper, either high or low values on the latent trait are associated with high person-fit. But, importantly, the models also suggest that in most instances the latent trait only explains between 2 and 4 percent of the variance in person-fit. The results of these analyses can be obtained from the authors upon request.

actual changes in the family composition (birth, marriage, divorce, death) and job changes as well as temporal changes in the individuals' position on the latent concepts as indicators of traitedness.

We find the expected higher levels of person-fit in persons who are emotionally highly stable. Also, we find lower levels of person reliability in persons currently undergoing critical life events and those who changed their self-assessment on the concepts analyzed in this study.

Demographics

Some studies discuss gender differences in person reliability and find either unsystematic or spurious relationships (Schmitt et al. 1999; Woods, Oltmanns, and Turkheimer 2008). We thus also control for gender and other demographics, such as age, region, and migration status, in the multivariate models. Table 3 reports an inverted U-shaped association between age and person reliability and some minor differences by gender, East and West Germans, and migration status, the latter possibly reflecting different levels of language proficiency.

Fieldwork

Changes in the fieldwork procedures of SOEP, such as the stepwise replacement of PAPI by CAPI, and changes in the average duration of interviews may spuriously suggest changes in person reliability. We therefore consider the mode of data collection, the duration of the interview, the position of instruments in the questionnaire, and changes in the interviewer. Indeed, we find differences in person reliability across all fieldwork characteristics. Due to substantial self-selection into the mode of data collection and the duration of the interview, positive effects of short and self-administered interviews should probably not be interpreted as reflecting causal effects. This is because especially those respondents who experience difficulties in the questionnaire will take more time and will make use of the interviewer assistance in CAPI and PAPI (reference) mode, while those who are very familiar with the

questionnaire will take less time and may choose to complete the questionnaire without assistance from the interviewer (SAQ). In line with previous research, we find lower levels of data quality in the middle of the questionnaire, which probably reflects changing levels of fatigue during the interview.

All in all, we find a number of expected effects of third variables on person-fit statistics and we use very different estimation strategies, but the positive and statistically significant effect of survey experience on person reliability turns out to be highly robust. This robustness also exists if we analyze person reliability in each of the 17 instruments individually (Table 2 for an overview).

Conclusions

The analysis of this paper provides evidence that the repeated interviewing in longitudinal surveys increases the reliability of individual survey responses with every wave of a panel. This not only improves the statistical efficiency of data emanating from panel compared to cross-sectional surveys, but it also diminishes problems of attenuation bias (see Online Appendix B for a detailed empirical example). The increase in reliability with only four years of annual survey experience is comparable to the increase in reliability between primary and tertiary education.

More than one million observations of person reliability estimates in almost 50,000 SOEP respondents who entered the ongoing survey at very different points in time allow us to rigorously test the effect of increasing survey experience controlling for aging effects, period effects, and many other fieldwork characteristics and person-related factors that may generate spurious relationships between survey experience and person reliability. Moreover, the panel design of our data supports the estimation of person-fixed-effects models and we also use mixed models to decompose variation in person reliability between observation, respondent,

and interviewer. In all these model specifications, we find a highly robust positive effect of survey experience on person reliability.

One may object that reliability of answers may also emerge from the tendency of respondents to provide consistent and non-contradictory answers. Waterton and Lievesley (1989) refer to this phenomenon as "freezing" of attitudes. In this event, the estimated person reliability of answers clearly is not indicative of data quality. In the present analysis, the time gap between surveying single multi-item constructs is between 1 and 7 years. In fact, in 11 out of 17 instruments, the gap is four years or more (see Table 2). Previous research suggests that respondents may recall that previous interviews included questions on, for instance, personality traits, but that they do not recall their exact answers. According to Van Meurs and Saris (1990), in the context of MTMM studies and in many items, a 25-minute gap between replications is sufficient to obtain independent measures. This is why we think that the desire to appear consistent is not the primary causal process that leads to increased person reliability.

Cognitive theories of survey response offer manifold reasons to expect that survey experience improves data quality. Clearly, more research is needed to unequivocally identify the exact causal processes of this effect: whether survey experience diminishes ambiguities of comprehension of survey questions, whether it facilitates the retrieval of relevant information, whether it reduces the need for heuristics of judgment, or whether it makes haphazard misreporting less likely. This paper provides first evidence that the increased person reliability is presumably *not* due to the familiarly between respondent and interviewer. Rather, we find that the improvement is a function of both general experience in the SOEP as well as instrument-specific experience (Sturgis et al. 2009). The positive effects of general survey experience suggest in particular that beyond learning processes relating to single items, such as improved comprehension of a survey question that is repeated again in subsequent years, experience in answering one question may also help to improve respondents' ability to answer

new survey questions as well. Understanding the survey task as a whole thus seems to be a factor in increased person reliability in panel surveys that is independent of the type of questions asked.

References

Bailar, Barbara A. 1989. "Information Needs, Surveys, and Measurement Errors." In *Panel Surveys*, eds. Daniel Kasprzyk, Greg Duncan, Graham Kalton, and M.P. Singh, 1-24. New York: Wiley.

Basso, Michael R., Natasha Lowery, Courtney Ghormley, and Robert A. Bornstein. 2001. "Practice Effects on the Wisconsin Card Sorting Test–64 Card Version Across 12 Months." *The Clinical Neuropsychologist* 15:471-78.

Binswanger, Johannes, Daniel Schunk, and Vera Toepoel. 2013. "Panel Conditioning in Difficult Attitudinal Questions." *Public Opinion Quarterly* 77:783-97.

Bradburn, Norman. 1978. "Respondent Burden." Proceedings of the Survey Research Methods Section of the American Statistical Association, 35-40.

Brown, Ryan P. 2003. "Measuring Individual Differences in the Tendency to Forgive: Construct Validity and Links with Depression." *Personality and Social Psychology Bulletin* 29:759-71.

Campbell, Donald T., and Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.

Cantor, David. 2008. “A Review and Summary of Studies on Panel Conditioning.” In *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, ed. Menard S. Burlington, 123-238. CITY, MA: Academic Press.

Chandon, Pierre, Vicki G. Morwitz, and Werner J. Reinartz. 2005. "Do Intentions Really Predict Behavior? Self-generated Validity Effects in Survey Research." *Journal of Marketing* 69 (2):1-14.

Conijn, Judith M. 2013. *Detecting and Explaining Person Misfit in Non-Cognitive Measurement*. Doctoral Thesis, Tilburg University.

Dagohoy, Anna V. T. 2005. *Person Fit for Tests with Polytomous Responses*. Doctoral Thesis, University of Twente.

De Leeuw, Edith D., and Joop J. Hox. 1994. "Are Inconsistent Respondents Consistently Inconsistent? A Study of Several Nonparametric Person Fit Indices." In *Measurement Problems in the Social Sciences*, eds. Joop J Hox and Wim Jansen, 67-88. Amsterdam: SISWO.

Drasgow, Fritz, Michael V. Levine, and Esther A. Williams. 1985. "Appropriateness Measurement With Polychotomous Item Response Models and Standardized Indices." *British Journal of Mathematical and Statistical Psychology* 38:67-86.

Fazio, Russell H. 1989. "On the Power and Functionality of Attitudes." In *Attitude Structure and Function*, eds. Anthony R. Pratkanis, Steven J. Breckler, and Anthony G. Greenwald, 153-79. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Feldman, Jack M., and John G. Lynch. 1988. "Self-generated Validity and Other Effects of Measurement on Belief, Attitude, Intention, and Behavior." *Journal of Applied Psychology* 73:421-35.

Fowler, Floyd J. 1992. "How Unclear Terms Affect Survey Data." *Public Opinion Quarterly* 56:218-31.

Fowler, Floyd J. 1995. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks: Sage.

Holbrook, Allyson L., Melanie C. Green, and Jon A. Krosnick. 2003. "Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias." *Public Opinion Quarterly* 67:79-125.

Holt, D. Tim 1989. "Panel Conditioning: Discussion." In *Panel Surveys*, eds. Daniel Kasprzyk, Greg Duncan, Graham Kalton, and M.P. Singh, 340-47. New York: Wiley.

Jagodzinski, Wolfgang, Steffen M. Kühnel, and Peter Schmidt. 1987. "Is there a "Socratic Effect" in Nonexperimental Panel Studies?" *Sociological Methods & Research* 15:259-302.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect theory: An Analysis of Decision under Risk." *Econometrica* 47:263-92.

Karabatsos, George. 2003. "Comparing the Aberrant Response Detection Performance of Thirty-six Person-fit Statistics." *Applied Measurement in Education* 16:277-98.

Knäuper, Barbel, Robert F. Belli, Daniel H. Hill, and Anna R. Herzog. 1997. "Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality." *Journal of Official Statistics* 13:181-99.

Kraut, Robert E., and John B. McConahay. 1973. "How Being Interviewed Affects Voting: An Experiment." *Public Opinion Quarterly* 37:398-406.

Kroh, Martin. 2013. "Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) (1984 until 2012)." DIW Berlin.

Krosnick, Jon A. 1999. "Survey Research." *Annual Review of Psychology* 50:537-67.

Krosnick, Jon A., and Duane F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response-order Effects in Survey Measurement." *Public Opinion Quarterly* 51:201-19.

Krosnick, Jon A., Sowmya Narayan, and Wendy R. Smith. 1996. "Satisficing in Surveys: Initial Evidence." *New Directions for Evaluation* 1996 (70):29-44.

Lang, Frieder R., David Weiss, Andreas Stocker, and Bernhard von Rosenbladt. 2007. "Assessing Cognitive Capacities in Computer-Assisted Survey Research: Two Ultra-Short Tests of Intellectual Ability in the German Socio-Economic Panel." *Schmollers Jahrbuch* 127:183-91.

Lazarsfeld, Paul F. 1940. "Panel Studies." *Public Opinion Quarterly* 4:122-28.

Lazarsfeld, Paul F., and Marjorie Fiske. 1938. "The “Panel” as a New Tool for Measuring Opinion." *Public Opinion Quarterly* 2:596-612.

Levine, Michael V., and Donald B. Rubin. 1979. "Measuring the Appropriateness of Multiple-choice Test Scores." *Journal of Educational and Behavioral Statistics* 4:269-90.

Masters, Geoff N. 1982. "A Rasch Model for Partial Credit Scoring." *Psychometrika* 47:149-74.

McFarland, Lynn A., and Ann M. Ryan. 2000. "Variance in Faking across Noncognitive Measures." *Journal of Applied Psychology* 85:812-21.

Meijer, Rob R., and Klaas Sijtsma. 2001. "Methodology Review: Evaluating Person Fit." *Applied Psychological Measurement* 25:107-35.

Nancarrow, Clive, and Trixie Cartwright. 2007. "Online Access Panels and Tracking Tesearch. The Conditioning Issue." *International Journal of Market Research* 49:573-94.

Porst, Rolf, and Klaus Zeifang. 1987. "A Description of the German General Social Survey Test-retest Study and a Report on the Stabilities of the Sociodemographic Variables." *Sociological Methods & Research* 15:177-218.

Reise, Steven P., and Niels G. Waller. 1993. "Traitedness and the Assessment of Response Pattern Scalability." *Journal of Personality and Social Psychology* 65:143-51.

Richardson, Florence, and Edward S. Robinson. 1921. "Effects of Practice upon the Scores and Predictive Value of the Alpha Intelligence Examination." *Journal of Experimental Psychology* 4:300-17.

Richter, David, Maria Metzing, Michael Weinhardt, and Jürgen Schupp. 2013. "SOEP Scales Manual." DIW Berlin.

Schmitt, Neal, David Chan, Joshua M. Sacco, Lynn A. McFarland, and Danielle Jennings. 1999. "Correlates of Person Fit and Effect of Person Fit on Test Validity." *Applied Psychological Measurement* 23:41-53.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Content.* New York: Academic Press.

Schwarz, Norbert. 2007. "Cognitive Aspects of Survey Methodology." *Applied Cognitive Psychology* 21:277-87.

Sikkel, Dirk, and Adriaan Hoogendoorn. 2008. "Panel Surveys." In *International Handbook of Survey Methodology*, eds. Edith D. De Leeuw, Joop J. Hox, and Don A. Dillman, 479-99. New York: Lawrence Erlbaum.

Sturgis, Patrick, Nick Allum, and Ian Brunton-Smith. 2009. "Attitudes Over Time: The Psychology of Panel Conditioning." In *Methodology of Longitudinal Surveys*, ed. Peter Lynn, 113-26. Chichester, UK: John Wiley & Sons, Ltd.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Tellegen, Auke. 1988. "The Analysis of Consistency in Personality Assessment." *Journal of Personality* 56:621-63.

Toepoel, Vera, Marcel Das, and Arthur van Soest. 2009. "Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents." *Survey Research Methods* 3:73-80.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Traugott, Michael W., and John P. Katosh. 1979. "Response Validity in Surveys of Voting Behavior." *Public Opinion Quarterly* 43:359-77.

van Krimpen-Stoop, Edith M. L. A., and Rob R. Meijer. 2002. "Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items." *Applied Psychological Measurement* 26:164-80.

Van Meurs, A., and Willem E. Saris. 1990. “Memory Effects in MTMM Studies.” In *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*, ed. Willem E. Saris, and A. van Meurs, 134–46. Amsterdam: North Holland.

van Vaerenbergh, Yves, and Troy D. Thomas. 2013. "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies." *International Journal of Public Opinion Research* 25:195-217.

Wagner, Gert G., Joachim R. Frick, and Jürgen Schupp. 2007. "The German Socio-Economic Panel Study (SOEP) - Scope, Evolution and Enhancements." *Schmollers Jahrbuch* 127:139-69.

Warren, John R., and Andrew Halpern-Manners. 2012. "Panel Conditioning in Longitudinal Social Science Surveys." *Sociological Methods & Research* 41:491-534.

Waterton, Jennifer, and Denise Lievesley. 1989. "Evidence of Conditioning Effects in the British Social Attitudes Panel Survey." In *Panel Surveys*, eds. Daniel Kasprzyk, Greg Duncan, Graham Kalton, and M.P. Singh, 319-39. New York: Wiley.

Woods, Carol M., Thomas F. Oltmanns, and Eric Turkheimer. 2008. "Detection of Aberrant Responding on a Personality Scale in a Military Sample: An Application of Evaluating Person Fit with Two-level Logistic Regression." *Psychological Assessment* 20:159-68.

Zaller, John, and Stanley Feldman. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36:579-616.

Figure 1. Exemplary Response Patterns for "Tendency to Forgive

The answers of individuals [X] and [O] suggest about the same level of forgivingness in both individuals (θ_X =-0.70 and θ_O=-0.72). However, the first response pattern of [X] is associated with high levels of person-fit and the second response pattern of [O] with low fit (l_{zX}=1.4 and l_{zO}=-3.4). The predicted probability of selecting single response options is highest if their diffiulty (reported between boxes) matches the individual score on the latent trait. Little overlap between expected and observed answers implies an abbrarent response patterns and thus low person-fit. Table 1 reports negative discrimination parameters λ in the case of reversely formulated items. For reasons of clarity, the example of Figure 1 uses the original, reversed response options instead.

Figure 2 Distribution of Person-Reliability (SOEP, N=1,079,900)

Table 1 Two-Parameter Item Response Models of Multi-Item Constructs (SOEP, 1992—2013)

Trait	Item	Response Categories	Difficulty (min-max)			Discrimination (Ref. Item 1)
	i	m_i	δ_{ik}			λ_i
Openness	is original, comes up with new ideas	7	-4.8	-	3.3	1.0
	values artistic, aesthetic experiences	7	-2.6	-	2.4	0.7
	has an active imagination	7	-5.1	-	2.7	1.1
Conscient.	does a thorough job	7	-9.2	-	0.3	1.0
	tends to be lazy	7	-0.5	-	4.6	-0.3
	does things effectively and efficiently	7	-5.7	-	1.4	0.5
Extraversion	is communicative, talkative	7	-7.9	-	1.8	1.0
	is outgoing, sociable	7	-6.6	-	2.6	0.9
	is reserved	7	-3.0	-	3.1	-0.4
Neuroticism	worries a lot	7	-3.6	-	2.3	1.0
	gets nervous easily	7	-3.7	-	4.8	2.2
	is relaxed, handles stress well	7	-4.1	-	2.9	-1.2
Agreeableness	is sometimes rude to others	7	-1.4	-	4.1	-1.0
	has a forgiving nature	7	-5.2	-	1.4	1.0
	is considerate and kind to others	7	-9.9	-	2.2	3.4
Positive	returns favors	7	-7.1	-	-0.7	1.0
Reciprocity	helps those who help him/her	7	-8.7	-	1.1	2.1
	helps those who have helped him/her	7	-4.6	-	1.6	1.0
Negative	gets revenge for severe injustices	7	-2.9	-	5.5	1.0
Reciprocity	causes similar problems	7	-3.1	-	8.4	1.4
	insults those who insult him/her	7	-2.1	-	4.2	0.6
Tendency to	overcomes emotional violation fast	7	-2.5	-	3.6	1.0
Forgive	thinks about experien. injustice long	7	-3.7	-	1.9	-1.0
	tends to bear grudges	7	-2.1	-	4.0	-1.3
	tries to forget injustice done to him/her	7	-3.6	-	3.3	1.2
Physical Strain	current health	5	-3.5	-	4.9	-1.0
	state of health affects ascending stairs	3	-3.4	-	-0.6	1.0
	state of health affects tiring tasks	3	-3.1	-	0.0	1.2
	strong physical pain last 4 weeks	5	-6.0	-	0.9	1.2
	accomplished less due to physical prob.	5	-11.3	-	1.5	2.7
	limitations due to physical problems	5	-13.4	-	1.3	3.4
Mental Strain	pressed for time last 4 weeks	5	-3.4	-	1.9	1.0
	run-down, melancholy last 4 weeks	5	-5.7	-	2.0	2.9
	well-balanced last 4 weeks	5	-3.1	-	4.5	-2.0
	used lot of energy last 4 weeks	5	-3.4	-	3.6	1.7
	accomplished less due to emot. prob.	5	-17.8	-	-0.9	12.7
	less careful due to emotional problems	5	-13.7	-	-1.1	9.1
Well-being	freq. of being angry in last 4 weeks	5	-2.5	-	3.7	1.0
	freq. of being worried in last 4 weeks	5	-0.6	-	5.7	1.6
	freq. of being happy in last 4 weeks	5	-4.0	-	2.6	-0.7
	freq. of being sad in the last 4 weeks	5	-2.3	-	6.0	2.2
Work Effort	job-related burdens: high time pressure	2			0.6	1.0
	job-related burdens: freq. interruptions	2			0.1	0.6
	jobs-related burdens: amount of work	2			0.5	0.7
Trust	trusts people on the whole	4	-4.5	-	5.1	-1.0
	nowadays can't trust anyone	4	-3.9	-	3.4	0.9
	caution towards strangers	4	-0.3	-	4.4	0.4
	most people are exploitive/ fair	2			-0.2	0.2

Table 1 Two-Parameter Item Response Models of Multi-Item Constructs (SOEP, 1992—2013)

Trait	Item	Response Categories	Difficulty (min-max)			Discrimination (Ref. Item 1)
	i	m_i	δ_{ik}			λ_i
	most people are helpful	2			0.6	0.5
	freq. of lending friends pers. belong.	5	-3.5	-	1.6	-0.2
	freq. of lending friends money	5	-5.4	-	-0.1	-0.1
	freq. of leaving door unlocked	5	-3.0	-	-0.3	-0.1
External control	plans seldom work out	4	-3.6	-	1.9	1.0
beliefs 1	no one can escape fate	4	-1.6	-	2.7	0.9
	I get something because of luck	4	-3.4	-	1.6	0.8
	something unforeseen happens	4	-4.3	-	3.5	1.8
	the outcome is always different	4	-3.8	-	3.1	1.7
External control	my life's course depends on me	7	-1.2	-	4.7	1.0
beliefs 2	what you achieve depends on luck	7	-2.3	-	3.1	1.0
	others make the crucial decisions	7	-1.8	-	4.3	1.7
	doubt my abilities when problems arise	7	-1.9	-	4.0	1.3
	possib. are defined by social conditions	7	-3.5	-	2.5	0.8
	little control over my life	7	-1.9	-	5.8	2.5
Anomie	is confident about future	4	-2.2	-	2.9	-1.0
	feels lonely	4	-3.3	-	0.1	1.5
	does not enjoy work	4	-3.6	-	0.2	1.5
	is barely able to cope with things	4	-4.2	-	0.2	2.1
Risk Aversion	personal willingness to take risks	11	-3.7	-	6.7	1.0
	willingness to take risks while driving	11	-1.7	-	6.9	1.0
	will. to take risks in financial matters	11	-1.1	-	7.2	1.0
	will. to take risks in leisure and sports	11	-2.3	-	7.0	1.2
	willingness to take risks in occupation	11	-2.0	-	6.5	1.2
	willingness to take health risks	11	-1.7	-	6.6	1.0
	will. to take risks after win. the lottery	7	0.9	-	4.5	0.4

Table 2 Person Reliability by Multi-Item Constructs (SOEP, 1992—2013)

Trait	Number of Waves	Gap between Waves	Number of Observations (Person x Wave x Instr.)	Person Reliability (l_z)		Effect[1] of Experience
				Mean	SD	
Openness	3	4	60,035	0.39	0.94	
Conscientiousness	3	4	60,180	0.28	0.86	
Extraversion	3	4	60,373	0.39	0.91	+
Neuroticism	3	4	60,394	0.43	0.91	+
Agreeableness	3	4	60,393	0.32	0.90	+
Positive Reciprocity	2	5	39,668	0.28	0.83	+
Negative Reciprocity	2	5	39,570	0.36	1.09	+
Tendency to Forgive	1	single obs.	18,702	0.37	1.05	single obs.
Physical Strain	6	2	125,108	0.23	1.08	+
Mental Strain	6	2	125,531	0.22	1.15	+
Well-being	7	1	140,365	0.31	0.95	+
Work Effort	2	5	23,687	0.38	0.35	?
Trust	3	5	58,862	0.29	1.02	+
External control beliefs 1	3	1	35,837	0.29	1.08	+
External control beliefs 2	2	5	39,216	0.32	1.06	
Anomia	7	irregular	94,887	0.29	1.08	+
Risk Aversion	2	5	36,260	0.22	1.04	+
Median/Total	3	4	1,079,068	0.31	1.02	+

[1] Based on Panel Fixed Effects Specification Equivalent to Model (2) reported in Table 3.

Table 3 The Effect of Survey Experience on Person Reliability (SOEP, 1992—2013)

		(1) P-OLS	(2) FE	(3) MIX	(4) P-OLS	(5) FE	(6) MIX
Experience	*General Survey Experience in Years (ref=1 year)*						
	2	0.02**	0.04**	0.03*	0.03*	0.03**	0.05**
	3	0.03**	0.05**	0.04**	0.03*	0.04*	0.04*
	4	0.09**	0.11**	0.10**	0.07**	0.08**	0.10**
	5-9	0.07**	0.10**	0.08**	0.06**	0.06**	0.08**
	10+	0.08**	0.12**	0.10**	0.08**	0.08**	0.10**
	Familiarity with Interviewer in Years (ref=1 year)						
	2	0.01	0.00**	-0.00	0	0.01**	-0.02
	3	0.03**	0.02	0.02	0.02	0.03	-0.00
	4	0.02**	0.01**	0.01	0.01	0.02**	-0.01
	5-9	0.02**	0.02*	0.01	0.01	0.03	-0.01
	10+	0.02**	0.02**	-0.01	0.01	0.03*	-0.02
	Familiarity with Instrument in Years (ref=1 year)						
	2	0.05**	0.03**	0.04**	0.03**	0.03**	0.03**
	3	0.05**	0.02**	0.04**	0.04**	0.04**	0.04**
	4+	0.08**	0.05**	0.06**	0.07**	0.07**	0.07**
Cognitive Ability	*Education (ref=Primary)*						
	Secondary				0.04**		0.05**
	Tertiary				0.06**		0.08**
	Word Fluency Test (ref=Low Value)						
	Medium				0.05**		0.03**
	High				0.06**		0.04*
	Symbol Digit Test (ref=Low Value)						
	Medium				0.03**		0.01
	High				0.03*		0.01
Survey Effort	*Straightlining*				0.09**	0.04**	0.06**
	Conscientiousness (ref=Low Value)						
	Medium				-0.04**		-0.03**
	High				-0.20**		-0.19**
	Item Nonresponse				-0.95**	-0.26**	-0.55**
Response Behavior	*Prospective Panel Participation in Years (ref=1 year)*						
	2				0.02**	0.02**	0.02**
	3				0.01*	0.01**	0.01
	4				0.03**	0.03*	0.02**
	5-9				0.03**	0.02**	0.02**
	10+				0.06**	0.05**	0.05**
	Latent Trait (ref=Moderate Value)						
	Directional				0.05**	0.02**	0.03**
	Extreme				0.46**	0.55**	0.52**
Traitedness	*Neuroticism (ref=low)*						
	Medium				0.01**		0.01*
	High				-0.08**		-0.07**
	Family Change				-0.03**	-0.02**	-0.02**
	Job Change				-0.04**	-0.01**	-0.02**
	Change in Latent Trait (ref=Minor Change)						
	Some Change				-0.09**	-0.06**	-0.07**
	Large Change				-0.22**	-0.14**	-0.16**
	Single Observation				-0.10**	-0.06**	-0.08**
Demographics	*Female*				0.07**		0.06**
	Age				0.01**	0.02**	0.01**
	Age^2				-0.00**	-0.00**	-0.00**
	Migrant				-0.02**		-0.03**
	East German				0.02**		0

Fieldwork	*Mode of interview (ref=PAPI)*						
	SAQ				0.04**	0.03**	0.04**
	CAPI				-0.06**	-0.03**	-0.05**
	MAIL/ELSE				0.01*	0.04**	-0.01
	Duration of interview (ref=Short)						
	Medium				-0.01*	-0.01**	-0.00
	Long				-0.03**	-0.01	-0.02**
	Question Number (ref=1-83)						
	86-121				-0.06**	-0.06**	-0.06**
	124-152				-0.03**	-0.03**	-0.03**
	Change in Interviewer				-0.00	0.01**	
Intercept		0.13**	0.14**	0.15**	-0.12**	-0.38**	-0.14**
Observations		1,079,068	1,079,068	807,648	1,079,068	1,079,068	807,648
Respondents		49,522	49,522	42,430	49,522	49,522	42,430
Interviewer				1,296			1,296
Accounted Variance Respondents			23.3%	8.0%		24.9%	7.3%
Accounted Variance Interviewer				2.4%			2.1%

Note.: *p<..05; **p<.01. All Models contain fixed effects for refreshment samples, instruments, and waves.

Table 4 Descriptive Statistics of Explanatory Variables (SOEP, 1992—2013)

			Mean / Proportion	Std. Dev.	Range	N Persons x Waves
Experience	General Survey Experience in Years		9.68	7.05	1-30	324,033
	Familiarity with Interviewer in Y.		5.78	5.07	1-30	324,033
	Familiarity with Instrument in Y.		2.07	1.45	1-7	324,033
Cognitive Ability	Education	Primary	0.43		0-1	315,721
		Secondary	0.38		0-1	
		Tertiary	0.20		0-1	
	Symbol Digit Test		0.47	0.16	0-1	80,134
	Word Fluency Test		0.26	0.11	0-1	53,485
Survey Effort	Conscientiousness		0.03	2.65	-10.39-4.30	279,973
	Item Nonresponse		0.02	0.03	0-0.53	324,033
	Straightlining		0.23	0.42	0-1	324,025
Response Behavior	Latent Trait	Moderate Value	0.72		0-1	324,033
		Directional Value	0.20		0-1	
		Extreme Value	0.09		0-1	
	Prospective Participation in Years		6.58	5.26	1-22	324,033
Traitedness	Neuroticism		0.02	0.80	-2.29-2.48	280,611
	Family Change		0.15	0.36	0-1	324,033
	Job Change		0.14	0.35	0-1	324,033
	Latent Trait	Minor Change	0.34		0-1	324,033
		Some Change	0.31		0-1	
		Large Change	0.16		0-1	
		Single Observat.	0.20		0-1	
Demographics	Age		47.76	17.62	16-103	324,033
	Female		0.52	0.50	0-1	324,033
	East Germany		0.27	0.45	0-1	324,033
	Migrant		0.18	0.39	0-1	323,181
Fieldwork	Mode	PAPI	0.28			324,033
		PAPI/SAQ	0.04			
		SAQ	0.29			
		CAPI	0.27			
		MAIL	0.13			
	Duration of interview in minutes		33.45	14.18	1-355	249,856
	Change in Interviewer		0.06	0.24	0-1	324,033
	Question Number		88.98	50.95	1-152	324,033