Journal of Survey Statistics and Methodology (2018) **6**, 306–334

ESTIMATING SURVEY QUESTIONNAIRE PROFILES FOR MEASUREMENT ERROR RISK

BARRY SCHOUTEN*
FRANK BAIS
VERA TOEPOEL

Surveys differ in their topics, language, style, and design, and consequently, in their sensitivity to measurement error. Survey literature presents a range of characteristics of survey items that are assumed to be related to the magnitude and frequency of measurement error. In terms of questionnaire design and testing, it would be very useful to have a questionnaire profile that is a summary of the characteristics of the items contained in a questionnaire. This holds especially true in the context of multi-mode surveys where the detection of measurement error is crucial. The questionnaire profiles may be derived from scores that coders assign to the items in a questionnaire. Given that agreement among coders may be relatively low, as we observe, the number of coders must be large to ensure sufficient precision of the profiles. For multiple surveys, the coding workload may then become infeasible. In this paper, we propose methodology for the estimation of questionnaire profiles when a pool of coders is randomly allocated to a series of surveys. The methodology is based on multiple imputation and applied to 11 general purpose surveys in the Netherlands.

Barry Schouten is Professor and holds a PhD in mathematics with Statistics Netherlands, Department of Methodology and IT, Henri Faasdreef 312, 2492JP, The Hague, The Netherlands, Barry Schouten, Frank Bais holds a master in Political Sciences. He is currently PhD candidate and Vera Toepoel is Assistent Professor and holds a PhD in Social Sciences are with Utrecht University, Faculty of Social and Behavioral Sciences, Department of Methodology and Statistics, Padualaan 14, 3584CH, Utrecht, The Netherlands.

Acknowledgements: The authors would like to thank Corrie Vis, Edith de Leeuw, Joop Hox, Judit Arends, Mattijn Morren, Natalia Kieruj, Peter Lugtig, Rachel Vis, and Salima Douhou for their help in preparing a coding scheme and in coding the LISS panel core studies.

*Address correspondence to Barry Schouten Statistics Netherlands, Department of Methodology and IT, and Utrecht University, Faculty of Social and Behavioral Sciences, Department of Methodology and Statistics, The Netherlands; E-mail: bstn@cbs.nl.

doi: 10.1093/jssam/smx027 Advance access publication 25 October 2017

KEYWORDS: Data quality; Measurement error; Mode effect; Multiple imputation; Question characteristics.

1. INTRODUCTION

Measurement error is widely studied in the survey methodology literature, e.g., Alwin and Krosnick (1991), Biemer, Groves, Lyberg, Mathiowetz, and Sudman (1991), Fowler (1995) and Tourangeau, Rips, and Rasinski (2000), and it is known to be an error that is difficult to measure and predict. Various authors have attempted to develop methodology to predict measurement error from the characteristics of a survey item. The most well-known attempts are the Question Understanding Aid (QUAID) by Graesser, Cai, Louwerse, and Daniel (2006) and the Survey Quality Predictor (SQP) by Saris and Gallhofer (2007). More recent attempts to construct lists of predictive item characteristics are Campanelli, Nicolaas, Jäckle, Lynn, Hope, et al. (2011), Beukenhorst, Buelens, Engelen, Van der Laan, Meertens et al. (2013) and Bais, Schouten, Lugtig, Toepoel, Arends-Toth (2015). In this paper, we define and estimate questionnaire profiles, which are summaries of item characteristics over the items in a survey. In order to do so, we employ the scores from a set of coders that independently worked on a range of surveys.

The motivation for the questionnaire profiles comes from the urgent need for a relatively cheap and quick assessment of the overall measurement error risk of a survey, especially in the early (re)design stages of a survey. We are specifically motivated by multi-mode survey (re)designs that are often driven by cost constraints. Budget and time pressure may bring extensive, cognitive questionnaire testing, and/or costly experimental studies to assess measurement effects between alternative designs into question. In order to decide to do such testing and/or experimentation, an informative but preliminary assessment of measurement error risk is imperative. Questionnaire profiles are, however, not a substitute for in-depth cognitive questionnaire testing. They may form a criterion for including testing in the survey (re)design and may function as a starting point for such tests. Additionally, they may form the incentive to do experimentation and to reserve more time to (re)design a survey. Hence, questionnaire profiles are foremost tools for survey coordinators and management to make decisions about the various (re)design stages, although they also may contain valuable information for questionnaire designers. Obviously, the preparation of the profiles themselves is an investment in time and budget; they should be viewed as part of a total quality control toolbox and are especially suited for repeated, larger surveys.

A questionnaire profile summarizes the frequencies of occurrence of a predefined set of relevant item characteristics over the items in a survey. What is deemed relevant depends on the context. The motivation for the present study comes from multi-mode survey designs, where mode-specific measurement bias can be unexpectedly high, may slow down redesigns and hamper publication. An example of large biases is given by Schouten, Brakel, Buelens, Laan, and Klausch (2013). For this reason, we focus on the most relevant characteristics for mode effects: difficult language in the question or answer categories, the question asks for sensitive information, the question is sensitive to strong emotions, the question is non-central (asks for knowledge that lies outside daily life), and the question may be presumed to be a filter question. For a discussion of these characteristics, see Van der Vaart, Van der Zouwen, and Dijkstra (1995), Tourangeau and Yan (2007), Kreuter, Presser, and Tourangeau (2008), Campanelli, Nicolaas, Jäckle, Lynn, Hope, et al. (2011) and Eckman, Kreuter, Kirchner, Jäckle, Tourangeau, et al. (2014). However, the methodology presented here is not specific to the selection of item characteristics. Nor is it specific to the purpose for which the item characteristics are used. Recent uses of characteristics are, for example, the explanation of survey response times (see Yan and Tourangeau, 2008; Couper and Kreuter, 2013; and Olson and Smyth, 2015), which are indirectly linked to measurement error. The important message from this paper is that care is needed in estimating and employing the occurrence of the characteristics.

The coding of item characteristics is very similar to questionnaire expert reviews (see Presser and Blair, 1994 and Olson, 2010). Expert reviews precede pre-testing and may signal items with a risk of low data quality. Olson (2010) concludes, based on validation data, that indeed experts are able to detect items with higher item non-response and lower reporting accuracy. The difference to expert reviews is that we do not necessarily assume that coders are questionnaire experts; it is sufficient that they received some training and understanding in the item characteristics.

In line with Olson (2010), Bais, Schouten, Lugtig, Toepoel, Arends-Toth, et al. (2015) show that inter-coder agreement for item characteristics can be low, even for motivated, trained, and experienced coders. They conclude that disagreement can only be resolved by restrictive definitions of the characteristics or by very time-consuming item-by-item decisions to reach a consensus. How to summarize scores on characteristics into profiles is, therefore, not as straightforward as it may seem. One may simply estimate the average number of coders that scored an item as having the property and then take the mean of these averages over all items. By doing so, a lot of information about the coding (and, hence, the characteristics of the questionnaire profiles) is lost. Instead, we propose to construct a probability distribution for an arbitrary item of the survey to have the characteristic of interest. For each pair of survey and characteristic, the distribution is estimated by the empirical distribution based on the coder scores. The set of distributions per survey over all characteristics we term the questionnaire profile. We believe this profile to be more useful than simple means because all information is maintained but structured.

The precision of the resulting profile depends on the number of coders; the more coders, the more precise the estimated frequencies. Since the list of relevant item characteristics may be long and since coding is a time-consuming, and, consequently, costly exercise, it is usually infeasible to let a large number of coders work on all surveys. Hence, it becomes attractive to construct efficient coding and imputation schemes. It is important to stress that the coding of survey items concerns the wording and format of questions and answer categories, but not the actual answers given by the survey respondents.

There is a vast literature on inter-coder agreement, (see Cohen, 1960; Fleiss, 1971; and Shoukri, 2010), and various measures have been developed to evaluate agreement. The most well known is Cohen's Kappa and variants of this measure. This literature focuses on reliability of coding. When applied to our setting, it assumes that the same coder may give different scores for the same item when replicated at different times and in different circumstances. Here, we assume that coders worked conscientiously, and reliability is a negligible problem. The focus is on the systematic differences between coders, i.e., the validity of the scores. Given the findings of Bais et al. (2015), we believe that the systematic differences between coders dominate the random differences. For this reason, we do not consider the more traditional measures of inter-coder agreement.

As a useful by-product of the study, we give the questionnaire profiles of eleven multi-purpose surveys in the Netherlands scored by a group of eight coders. These surveys are conducted in a comparable form in many countries.

This paper reads as follows: In section 2, we define questionnaire profiles, and we propose an estimation strategy based on randomly allocated coders. In section 3, we apply the estimation strategy to the eleven general purpose surveys. In section 4, we end with a discussion.

2. ESTIMATING QUESTIONNAIRE PROFILES

In this section, we introduce a number of item characteristics, define questionnaire profiles based on these characteristics, and construct an estimation strategy.

2.1 Item Characteristics and Questionnaire Profiles

In Bais et al. (2015), an extensive list of item characteristics is presented. This list is derived from Saris and Gallhofer (2007), Campanelli et al. (2011) and Beukenhorst, Buelens, Engelen, Van der Laan, Meertens, et al. (2013). In this paper, we consider a subset of six characteristics that are taken from this list:

(1) Difficult language in question: the question contains one or more difficult words or a complicated sentence structure;
(2) Difficult language in answer: the answer categories contain one or more difficult words or requires a complicated cognitive action (e.g., sliding bars or abstract visual representation);
(3) Non-centrality: the question asks for knowledge or experience that lies outside daily life of the average respondent;

(4) Sensitive to emotions: the question may arouse negative emotions like anger, distress, sorrow or despair;
(5) Sensitive information: the question asks for information that is viewed as sensitive by the average respondent;
(6) Presumed filter question: the average respondent believes that the question is a filter question and some of the answer categories will avoid follow-up questions;

We view the selected item characteristics as the most influential on measurement error and, more specifically, as the most relevant to mode effects. We present these characteristics in this section, however, only to fix thoughts in the following. The methodology set out in this paper can be applied to any set of item characteristics.

Suppose one would like to code all items in a series of S surveys on a given item characteristic by human coders. A group of M coders is randomly assigned to the surveys, and each survey gets assigned A coders. In other words, each coder gets on average AS/M surveys to work on. First, we assume that $A = M$ and all coders do all surveys.

Let $C_{s,i,m}$ be the 0–1 score ($1 =$ characteristic present, $0 =$ absent) of coder m on item i in survey s for a certain item characteristic. The surveys are labeled $s = 1, 2, \ldots, S$, the coders are labeled $m = 1, 2, \ldots, M$, and let the items within surveys are labeled $i = 1, 2, \ldots, I_s$. Let $I = \sum_{s=1}^{S} I_s$ be the total number of items. We suppress an index for the item characteristic in order to avoid an overly complex notation.

We view coders as selected from a super-population of coders, i.e., there is an underlying $p_{s,i}$ of interest, which may be viewed as the average item characteristic probability for item i in survey s over all possible coders. The $p_{s,i}$ and their average over the items in a survey p_s, i.e., $p_s = \frac{1}{I_s}\sum_{i=1}^{I_s} p_{s,i}$, are parameters of interest. We do not model the selection of items and the clustering of items within surveys, but assume these as given.

The item characteristic probabilities $\{p_{s,i}\}_{1 \leq s \leq S, 1 \leq i \leq I_s}$ are assumed to be drawn independently from a distribution G with support $[0, 1]$. Conditional on the item characteristic probability $p_{s,i}$, the coder item characteristic probabilities $\{p_{s,i,m}\}_{1 \leq m \leq M}$ are assumed to be drawn independently from a distribution $F_{s,i}$ with expected value $p_{s,i}$. Item characteristic probabilities for the same coder, say $p_{s,i,m}$ and $p_{s,j,m}$, are allowed to be dependent due to coder effects. Since we have a given set of surveys and items, we do not further parameterize or attempt to explicitly model the coder item characteristic probability distributions, but rather, we resort to empirical distributions.

Given these assumptions, the scores for a given set of items and coders $\{C_{s,i,m}\}_{1 \leq s \leq S, 1 \leq i \leq I_s, 1 \leq m \leq M}$ are independent and follow Bernoulli distributions with parameters $p_{s,i,m}$. It can then be shown by exchangeability of the item scores that for any $M > 1$, for any pair of survey items (s, i) and $(\tilde{s}, j)$, and for any vector $(c_1, c_2, \ldots, c_M) \in \{0, 1\}^M$,

$$
\begin{aligned}
&P[C_{s,i,m} = c_m | C_{s,i,1} = c_1, \ldots, C_{s,i,m-1} = c_{m-1},\\
&\quad C_{s,i,m+1} = c_{m+1}, \ldots, C_{s,i,M} = c_M] = P[C_{\tilde{s},j,m} = c_m |\\
&\quad C_{\tilde{s},j,1} = c_1, \ldots, C_{\tilde{s},j,m-1} = c_{m-1}, C_{\tilde{s},j,m+1} = c_{m+1}, \ldots, C_{\tilde{s},j,M} = c_M]. \qquad (1)
\end{aligned}
$$

Expression (1) directly allows for imputation schemes.

The set of p_s over the multiple item characteristics may be viewed as a profile of a questionnaire. However, the p_s do not express the amount of coder (dis)agreement, i.e., the variability in coder probabilities. Two surveys may have the same average p_s over their items, but may differ strongly in terms of coder consensus. This difference is important in judging if and how survey designers should deal with measurement error risk. For this reason, we include coder variability. We let $f_p(x),\ x \in [0,1]$, be the probability density function for distribution F_p. Now, we define the survey average

$$P_s(x) = \frac{1}{I_s} \sum\nolimits_{i=1}^{I_s} f_{p_{s,i}}(x), \qquad (2)$$

as the questionnaire profile for an item characteristic. We have that $\int_0^1 P_s(x)dx = 1$ and $P_s(x)$ may be interpreted as the relative proportion of items in survey s that has the item characteristic according to a fraction x of the coders. If coders would fully agree, then $P_s(x) = 0$ for $0 < x < 1$.

The set of functions P_s over all item characteristics we call the full questionnaire profile or simply the questionnaire profile.

The $p_{s,i}$, p_s and $P_s(x)$ are unknown, and they are to be estimated. The obvious estimators are

$$\hat{p}_{s,i} = \frac{1}{M} \sum_{m=1}^{M} C_{s,i,m}, \qquad (3)$$

$$\hat{p}_s = \frac{1}{I_s M} \sum_{m=1}^{M} \sum_{i=1}^{I_s} C_{s,i,m}, \qquad (4)$$

and the empirical density function for (3) is

$$\hat{P}_s(x) = M \frac{1}{I_s} \sum\nolimits_{i=1}^{I_s} A_{s,i}(x), \qquad (5)$$

where $A_{s,i}(x)$ is the observed 0–1 indicator for the event that a fraction x of the coders scored the characteristic, i.e., that $\frac{1}{M} \sum_{m=1}^{M} C_{s,i,m} = x \in \{0, \frac{1}{M}, \frac{2}{M}, \ldots, 1\}$. For $x \notin \{0, \frac{1}{M}, \frac{2}{M}, \ldots, 1\}$, we simply interpolate. The multiplication by M in (5) results from the bin size $= \frac{1}{M}$.

Although, we are not specifically interested in the item characteristic probabilities of individual coders, we do estimate the coder average score, denoted by $\theta_m = \frac{1}{IM} \sum_{s=1}^{S} \sum_{i=1}^{I_s} p_{s,i,m}$, using

$$\hat{\theta}_m = \frac{1}{IM}\sum_{s=1}^{S}\sum_{i=1}^{I_s} C_{s,i,m}. \tag{6}$$

We do this, because in practice, the pool of available coders working on surveys will change only gradually over time, with new coders starting and former coders quitting at a low rate, so that we may need to monitor and maintain the individual average coder probabilities.

The standard errors for estimators (3), $\sigma_{s,i}$, (4), σ_s , (5), $\tau_s\,(x)$, and (6), σ_m are estimated using resampling methods. Since we observe only one score per coder per item, we cannot account for lack of coder reliability, due to $p_{s,i,m}$ between 0 or 1. In estimating standard errors for the $\hat{\theta}_m$, we ignore this variability.

In practice, it will usually be too costly to score the items of all surveys by all coders on all item characteristics. In the case study in section 3, the coders scored on average thirty-five items per hour on the set of characteristics, and coding all items would have cost roughly seventy hours per coder. However, as was concluded in Bais et al. (2015), the coder average scores may vary greatly and, as a consequence, multiple coders are needed to obtain a precise estimate of the item characteristic and survey probabilities. This leads to a trade-off between coder costs and coding precision. In the next section, we show how the various parameters of this section can be estimated using multiple imputation, when part of the coder scores are missing by design.

2.2 Multiple Imputation to Account for Missing Coder Scores

Instead of coding all surveys, assume the coders work only on a random subset of surveys, i.e., $A < M$. Assume, furthermore, that the coders differ in their maximal workload. Let S_m be the number of surveys that coder m can work on, i.e., $\sum_{m=1}^{M} S_m = AS$. Let $U_{m,s}$ be the 0–1 indicator for the allocation of survey s to coder m. We have that $\sum_{s=1}^{S} U_{m,s} = S_m$ and $P[U_{m,s} = 1] = S_m/S$.

We use multiple imputation to fill out the missing scores of coders and to estimate the questionnaire profiles. As an important by-product, we estimate the standard errors following from the missing item scores and the standard errors following from the selection of coders.

The algorithm is

(1) Construct an imputation scheme for the missing surveys;
(2) Repeat B times the following steps:
    (a) Perform a (random) imputation given the scheme of step 1;
    (b) Based on the imputed data set, estimate the item characteristic probability, $\hat{p}^b_{s,i}$, the survey probability, $\hat{p}^b_s$, the coder average score, $\hat{\theta}^b_m$, and the questionnaire profile, $\hat{P}^b_s(x)$.

(c) Based on the imputed data set, estimate the standard errors for the item probabilities, $\hat{\sigma}^{b}_{s,i}$, using (5), and for the survey probabilities, $\hat{\sigma}^{b}_{s}$, and the questionnaire profiles, $\hat{\tau}^{b}_{s}(x)$, using the bootstrap;

(3) Estimate the mean of the item characteristic probabilities, $\hat{p}_{s,i} = \sum_{b=1}^{B} \hat{p}^{b}_{s,i}$, survey probabilities, $\hat{p}_{s} = \sum_{b=1}^{B} \hat{p}^{b}_{s}$, questionnaire profiles, $\hat{P}_{s}(x) = \sum_{b=1}^{B} \hat{P}^{b}_{s}(x)$, and coder average scores, $\hat{\theta}_{m} = \sum_{b=1}^{B} \hat{\theta}^{b}_{m}$;

(4) Estimate the mean of the standard errors of the item characteristic probabilities, $\hat{\sigma}_{s,i}^{\;W} = \sum_{b=1}^{B} \hat{\sigma}^{b}_{s,i}$, the survey probabilities, $\hat{\sigma}^{W}_{s} = \sum_{b=1}^{B} \hat{\sigma}^{b}_{s}$, and the questionnaire profiles, $\hat{\tau}^{W}_{s}(x) = \sum_{b=1}^{B} \hat{\tau}^{b}_{s}(x)$;

(5) Estimate the standard deviation of the item characteristic probabilities, $\hat{\sigma}_{s,i}^{\;B} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{p}^{b}_{s,i} - \hat{p}_{s,i})^{2}}$, the survey probabilities, $\hat{\sigma}_{s}^{\;B} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{p}^{b}_{s} - \hat{p}_{s})^{2}}$, the questionnaire profiles, $\hat{\tau}_{s}^{\;B}(x) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{P}^{b}_{s}(x) - \hat{P}_{s}(x))^{2}}$, and the coder average scores, $\hat{\sigma}_{m}^{\;B} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{b}_{m} - \hat{\theta}_{m})^{2}}$;

(6) Estimate the total standard error using Rubin's rules, $\hat{\sigma}^{T}_{s,i} = \sqrt{(\hat{\sigma}_{s,i}^{\;W})^{2} + \left(1 + \frac{1}{M}\right)(\hat{\sigma}_{s,i}^{\;B})^{2}}$, $\hat{\sigma}^{T}_{s} = \sqrt{(\hat{\sigma}_{s}^{\;W})^{2} + \left(1 + \frac{1}{M}\right)(\hat{\sigma}_{s}^{\;B})^{2}}$, $\hat{\tau}^{T}_{s}(x) = \sqrt{(\hat{\tau}_{s}^{\;W}(x))^{2} + (1 + \frac{1}{M})(\hat{\tau}_{s}^{\;W}(x))^{2}}$, and $\hat{\sigma}_{m}^{\;T} = \hat{\sigma}_{m}^{\;B}$;

Some side remarks are in place: The variances of the estimators for complete data sets are called "within variances," whereas the variances over the imputed data sets are called "between variances." For this reason, the superscripts "W" and "B" are used in steps 4 and 5, respectively. The coder average scores have no standard error in a complete data set. Hence, standard errors arise only from the missing surveys; the within variances are zero by definition. A usual choice for the number of imputed data sets is $B = 10$. For the bootstrap, we use 1,000 replications per imputed data set.

The algorithm produces unbiased estimates under four conditions. First, the coders work independently from each other. Second, the coders score the items consistently, i.e., they score each item as isolated from the other items. Third, the surveys need to be allocated randomly to coders. Fourth, in the imputation, the matching property holds, i.e., for each missing score combination an observed score combination exists. The first two conditions are about the coders themselves. The third condition implies a missing-completely-at-random mechanism for the missing scores. The fourth condition implies that predictors exist for each missing score. The first three conditions are under control, and we assume they hold. However, the fourth condition does not hold necessarily and depends on the imputation scheme. For example, when the scores of nine coders are used to impute the scores of a tenth coder, then it is likely that part of the possible combinations of nine 0–1 scores did not occur in the data set. When such a combination occurs for a missing score on the tenth coder, then there is no observation that can be used to predict that missing score. Survey items are not randomly clustered within surveys, which increases the risk that the matching condition does not hold. In practice, therefore, parsimony is needed in the imputation scheme.

The imputation scheme is the most complicated part of the multiple imputation algorithm. The scheme describes the order in which surveys are imputed and the subset of coders that are used to predict the missing score. It is often not clear beforehand what is the most efficient scheme for a given allocation. The reduction in multiple imputation standard error depends on the correlation between the scores of coders, i.e., their mutual agreement, the amount of overlap in surveys between coders, and the amount of non-overlap between coders. When two coders always agree, then they form an ideal couple to impute each other's missing scores. When the agreement is the same between various pairs of coders, then the amount of overlap in items determines the order in which imputations are made; the more overlap the better. Finally, when agreement is the same and overlap is the same, it is the non-overlap that counts. As mentioned before, the matching condition warns against imputation using the scores of all available coders. In our case study, we use a maximum of three coders to impute scores of other coders.

Consider the example in table 1 with eight surveys and five coders. Per survey, three coders are assigned. Coders one through three can do six surveys, whereas coders four and five can only do three surveys. In total, twenty-four coded surveys are produced, and sixteen coded surveys are missing. Table 1 shows one possible realization of coder allocation. Given that coders one through three worked on six surveys, they show the largest overlap, and it is natural to impute their missing surveys first. Coders one and two and coders one and three worked on five surveys simultaneously. With no knowledge about their agreement, it is an option to impute the missing (coder, survey) cells (1,7), (2,1), (1,8) and (3,6). From there on, cell (2,8) can be imputed. Coders one, two, and three scored four surveys simultaneously (surveys two through five), and a second option is that pairs of coders are used to first impute (2,1) and (3,6) and then to proceed to the cells (1,7), (1,8) and (2,8). In either option, the (imputed) scores of the first three coders can then be used to impute surveys for coders four and five. However, depending on the agreement between coders, the optimal scheme could be different. For example, when coder four and five agreed on all items, then it may be more efficient to impute the missing cells (4,6) and (5,1) with each other's scores.

Another feature of the imputation scheme that has not been mentioned is that imputation needs to be performed per item characteristic. In general, the optimal scheme will be different for each characteristic, but different schemes may not be computationally attractive.

In section 3, we discuss an imputation scheme for a case study.

Table 1. Example of an Allocation of Five Coders to Eight Surveys. Gray Cells are Scores by the Coder. White Cells are Missing Survey Scores

Coder	Survey							
	1	2	3	4	5	6	7	8
1	▨	▨	▨	▨	▨	▨		
2		▨	▨	▨	▨	▨	▨	
3	▨	▨	▨	▨	▨			▨
4	▨						▨	▨
5						▨	▨	▨

3. A CASE STUDY: THE LISS PANEL CORE STUDIES AND DUTCH LABOR FORCE SURVEY

The estimation strategy of section 2.2 is applied to core studies of the Dutch LISS panel (Longitudinal Internet Studies for the Social Sciences) and the Dutch Labor Force Survey (LFS). We first describe the data and then present results.

3.1 The LISS Panel Data

The data that we will use comes from ten core study surveys that were all administered in the LISS panel of CentERdata, Tilburg University, 2008–2014. To these ten surveys, we added the Labor Force Survey, an ongoing monthly survey that is administered by Statistics Netherlands. LISS is a government-funded web panel based on a probability sample drawn by Statistics Netherlands. The panel was established in 2007 and has now been running for eight years. It is, generally, considered a high-quality panel because of the extensive recruiting and refreshment throughout the panel. The panel has roughly 8,000 panel members. Table 2 gives a short list of the surveys and topics contained in the surveys, and the number of items per survey. In total, the number of items is 2,470. From 2008 to 2014, all core study surveys have been administered annually, except for Assets (AS). The questionnaires of the last available wave were used for coding.

The coding of the surveys was prepared in four steps: First, a preliminary set of item characteristic definitions was made. Second, this set of definitions was applied by all coders to a small but broad set of items in a pilot study. Third the definitions were discussed and revised based on the pilot study findings. And fourth, they were applied to all items. The coders were two of the authors of the paper, two experts from Statistics Netherlands' cognitive questionnaire lab, three experts from CentERdata's questionnaire design department, and a mode effect expert from Utrecht University. In total, eight

Table 2. The Surveys in the Case Study

Label	Survey	Topics of the Content	I_s
AS	Assets	Assets, property, and investment;	50
FA	Family and Household	Household composition and family relations;	73
HE	Health	Health and well-being;	286
HO	Housing	Housing and household;	409
IN	Income	Employment, labor and retirement, income, social security, and welfare;	243
LFS	Labor Force Survey	Education, employment, and labor;	200
PE	Personality	Personality traits;	148
PO	Politics and Values	Politics, social attitudes, and values; attitudes towards surveys;	71
RE	Religion and Ethnicity	Religion, social stratification, and groupings;	396
SO	Social Integration and Leisure	Communication, language and media; leisure, recreation, and culture; social behavior, travel, and transport;	471
WO	Work and Schooling	Education, employment, labor, and retirement;	123

coders were available. To each survey, three coders were randomly allocated. However, their hours of availability were not equal, as is usually the case for coding exercises, so that the number of surveys per coder is very different; one coder did all surveys, one did ten surveys, one did five surveys, two did two surveys, and three did one survey. The coding exercise was time consuming since, apart from the six characteristics on which we focus attention in this paper, ten more characteristics were coded (Bais et al. 2015). A total of 2,470 items were coded on 16 characteristics. On average, thirty items could be coded per hour by one coder on all sixteen characteristics. For this reason, it was not possible to let all coders do all surveys; this avoided missing data. Table 3 presents the allocation to the surveys. We refer to Bais et al. (2015) for details.

In order to construct an imputation scheme, we looked at the correlations between the scores of coders that worked on the same surveys. These correlations turn out to be relatively low, in general. Table 4 shows the correlations for item characteristic "Sensitive information" for the pairs of coders that worked simultaneously on at least one survey. Given the low correlations and given the practicality of using the same scheme for all item characteristics, we decided to use the amount of overlap as the criterion to build the imputation scheme.

Table 3. The Allocation of Coders to the Case Study Surveys. The Light Gray Blocks are Omitted in the Sensitivity Analysis

Coder	AS	FA	HE	HO	IN	LFS	PE	PO	RE	SO	WO
1	░	▓	░	░	▓	▓	▓	▓	▓	▓	▓
2	▓			▓	▓				▓		▓
3	▓	░	▓	▓	░	▓	▓	▓		▓	▓
4		▓								▓	
5								▓	▓		
6						▓					
7			▓								
8							▓				

The imputation scheme we used is

(1) Impute the missing survey of coder 3 using the scores of coder 1;
(2) Impute the missing surveys of coder 2 using the scores of coders 1 and 3;
(3) Impute the missing surveys of all other coders using the scores of coders 1, 2 and 3;

For some of the coders, the majority of surveys were imputed, which has a strong impact on standard errors, as we will see in the next section. This points to the inefficient allocation of coders in table 4, which is the result of the strongly varying maximal workloads.

3.2 Results

We estimated the proportion of coders that would indicate an item to have a characteristic. We did this for all 2,470 items and for all six characteristics: difficult language in question (DLQ), difficult language in answer (DLA), noncentrality (CENT), sensitive to emotions (EMO), sensitive information (SENS), and presumed filter question (FILT). Furthermore, we aggregated the proportions over surveys and over coders, and we estimated questionnaire profiles per characteristic.

Table 5 gives the estimated survey probabilities $\hat{p}_s$ for the six item characteristics. For each estimate, two standard errors are given; the "within" standard error corresponds to a complete coder data set, and the total standard error also includes the imputation standard error. The "within" standard errors are much larger than the "between" standard errors, resulting from incomplete coder data. This points to a large uncertainty that is due to the coders. We will return to this conclusion when we discuss the coder probabilities. The surveys differ substantially in their probabilities. For instance, core study Assets (AS) scores highly on many of the characteristics and may be considered a survey that is

Table 4. Correlations between the Scores on Characteristic "Sensitive Information "for Pairs of Coders that Worked on At Least One Survey Simultaneously

	2	3	4	5	6	7	8
1	0.19	0.23	0.12	−0.02	0.47	0.17	0.14
2	–	0.31	NA	0.09	NA	NA	NA
3	–	–	0.22	0.23	0.59	0.63	0.02

Table 5. Estimated Probabilities Per Survey and Item Characteristic. Within Standard Errors and Total Standard Errors Within Brackets

	AS	FA	HE	HO	IN	LFS	PO	PE	RE	SO	WO
DLQ	32%	11%	13%	18%	23%	15%	18%	8%	21%	11%	21%
	(5.6)	(1.4)	(2.3)	(2.7)	(4.0)	(2.3)	(3.2)	(1.3)	(5.2)	(1.7)	(5.1)
	(6.2)	(1.5)	(2.4)	(2.8)	(4.1)	(2.5)	(3.3)	(1.3)	(5.3)	(1.8)	(5.1)
DLA	2%	7%	1%	4%	0%	7%	5%	2%	8%	2%	3%
	(1.2)	(1.9)	(0.3)	(1.0)	(0.3)	(2.1)	(1.6)	(0.6)	(2.1)	(0.5)	(0.7)
	(1.2)	(1.9)	(0.3)	(1.2)	(0.4)	(2.1)	(1.6)	(0.6)	(2.3)	(0.5)	(0.7)
CENT	33%	10%	13%	21%	28%	11%	21%	9%	23%	18%	20%
	(9.9)	(5.7)	(4.1)	(5.0)	(7.9)	(6.9)	(5.9)	(3.1)	(6.9)	(5.7)	(5.8)
	(10.1)	(5.7)	(4.2)	(5.0)	(8.0)	(7.1)	(6.0)	(3.2)	(7.1)	(5.7)	(5.8)
EMO	15%	15%	12%	11%	15%	10%	21%	18%	14%	10%	13%
	(6.8)	(5.4)	(5.9)	(5.6)	(5.8)	(5.9)	(7.9)	(5.8)	(5.6)	(5.6)	(5.5)
	(7.0)	(5.4)	(5.9)	(5.7)	(5.8)	(5.9)	(8.0)	(5.8)	(5.7)	(5.6)	(5.5)
SENS	58%	23%	34%	34%	44%	16%	35%	18%	42%	29%	27%
	(10.1)	(5.0)	(9.3)	(7.8)	(10.8)	(4.7)	(8.7)	(5.4)	(10.2)	(7.7)	(6.5)
	(10.4)	(5.1)	(9.4)	(7.9)	(10.8)	(4.8)	(8.7)	(5.4)	(10.4)	(7.7)	(6.5)
FILT	35%	29%	25%	25%	26%	30%	16%	11%	18%	30%	28%
	(4.5)	(4.0)	(4.1)	(4.9)	(3.4)	(5.9)	(3.3)	(3.6)	(3.0)	(4.3)	(4.5)
	(4.8)	(4.1)	(4.3)	(4.9)	(3.4)	(6.0)	(3.4)	(3.7)	(3.4)	(4.4)	(4.6)

susceptible to measurement error. Core study Personality (PE) has low scores and may be considered less prone to error. However, as mentioned, standard errors are large and confidence intervals are wide. Nevertheless, differences between surveys frequently test as significant because there is a strong covariance within coders over surveys. As a result, standard errors of differences between surveys are similar in magnitude to standard errors per survey.

Table 6 gives the estimated coder probabilities $\hat{\theta}_m$ for the six item characteristics. The standard errors are small. For coder one, they are zero by definition, since all surveys were allocated to this coder and we ignore coder reliability. The estimates confirm the large standard errors of table 5; there is a great

Table 6. Estimated Probabilities Per Coder and Item Characteristic. Standard Errors Within Brackets. Light Gray Cells Correspond to Coder-Survey Combinations Where Coders that Worked On One Survey did not Consider any Item to have the Characteristic

	1	2	3	4	5	6	7	8
DLQ	32%	8%	13%	12%	17%	13%	11%	22%
	(0.0)	(0.6)	(0.1)	(0.4)	(0.7)	(1.3)	(0.8)	(1.5)
DLA	7%	1%	4%	0.4%	3%	3%	3%	6%
	(0.0)	(0.2)	(0.1)	(0.1)	(0.3)	(0.3)	(0.2)	(1.2)
CENT	31%	5%	25%	0%	0%	27%	35%	16%
	(0.0)	(0.3)	(0.2)	(0.0)	(0.0)	(0.8)	(1.0)	(0.8)
EMO	7%	0%	19%	11%	0%	8%	52%	13%
	(0.0)	(0.1)	(0.1)	(0.6)	(0.0)	(0.5)	(0.9)	(0.4)
SENS	25%	5%	35%	9%	22%	59%	57%	27%
	(0.0)	(0.3)	(0.2)	(0.4)	(1.8)	(1.6)	(0.9)	(0.9)
FILT	20%	37%	25%	32%	34%	30%	27%	2%
	(0.0)	(0.7)	(0.1)	(0.7)	(1.6)	(1.8)	(1.5)	(0.3)

variability in scores between coders. It must be noted that the estimated probabilities may be biased for coders that did only a small number of surveys, as noted in section 2.2, despite the random allocation of coders to surveys. This fallacy appears when some combinations of scores over coders are absent in some of the surveys. For a number of coder-survey combinations, this occurred. These combinations are colored light gray in table 6. Remarkably, coders four and five did not score any of the items as being non-central, and coder five did not score any of the items as being sensitive to emotions as well. As a result, for these characteristics, these coders have an estimated probability equal to zero.

Figures 1 and 2 depict the estimated questionnaire profiles $\hat{P}_s(x)$ for all item characteristics for the LFS and core study Politics and values (PO), respectively. The profiles contain symmetric 95% confidence intervals (in gray) based on a normal approximation, but they cut off at zero. We realize that cutting off intervals is a crude way of avoiding negative values. However, non-normal asymmetric approximations are not straightforward, and we leave this to future research.

From figure 1, we see, for example, that 40% of the LFS items were estimated to be free of complex language in the question according to all coders, and 10% of the LFS items do have complex language in the question acoordng to half of the coders. For presumed filter question, these two percentages are both 20%, so there is much more variability among coders about this characteristic in the LFS. Remarkably, there is no characteristic where the estimated proportion of items that is scored by all coders is larger than zero, i.e., there

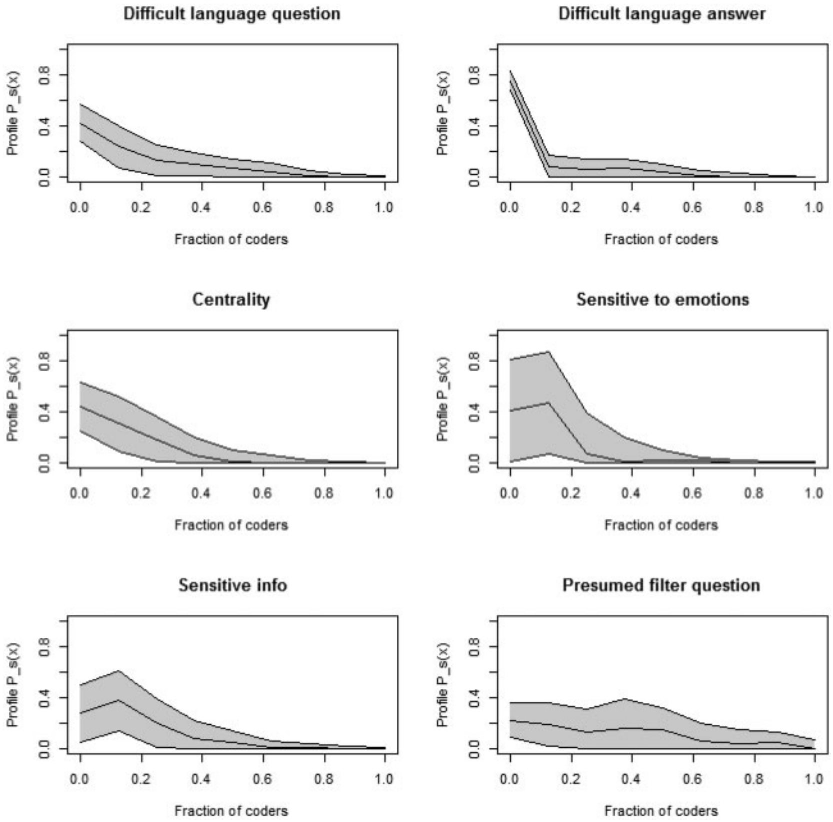

Figure 1. Questionnaire Profile for the Labor Force Survey (LFS). From left to right and from top to bottom: difficult language question, difficult language answer, centrality, sensitive to emotions, sensitive information and presumed filter question. Light gray cells represent 95% confidence intervals based on a normal approximation.

may never be a full consensus about items having one of the selected characteristics.

Despite the large standard errors, the profiles give a useful picture of the survey questionnaire. For the LFS, the fraction of items that a substantial number of coders, say 40%, would score as having the characteristic is only present for difficult language in question (DLQ) and presumed filter question (FILT). For core study Politics and values (PO), four of the characteristics show such fractions: difficult language in question (DLQ), non-centrality (CENT), sensitive to emotions (EMO), and sensitive information (SENS). Hence, the two surveys clearly have different profiles. Appendix A contains the profiles of the other nine surveys.

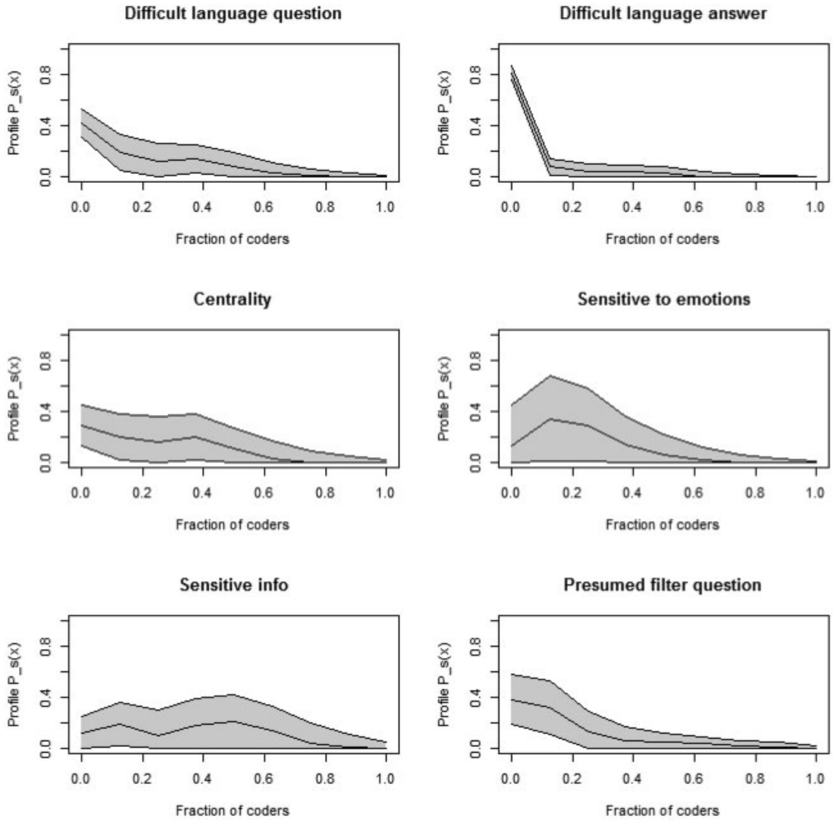

Figure 2. Questionnaire Profile for Core Study Politics and Values (PO). From left to right and from top to bottom: difficult language question, difficult language answer, centrality, sensitive to emotions, sensitive information and presumed filter question. Light gray cells represent 95**%** confidence intervals based on a normal approximation.

Table 7. Observed and Estimated Probabilities Per Item Characteristic

	DLQ	DLA	CENT	EMO	SENS	FILT
Observed	19%	4%	21%	12%	25%	26%
Estimated	16%	3%	17%	14%	30%	26%

Table 8. Differences in Estimated Survey Probabilities When Omitting Blocks for Coders One and Three. Underlined Differences Correspond to Values Outside 95% Confidence Intervals

	AS	*FA*	*HE*	*HO*	*IN*	*LFS*	*PO*	*PE*	*RE*	*SO*	*WO*
DLQ	−16%	1%	−7%	−10%	0%	−1%	0%	1%	0%	1%	0%
DLA	0%	0%	0%	−3%	0%	0%	0%	0%	0%	0%	0%
CENT	−7%	3%	−6%	−5%	−4%	0%	−4%	0%	−1%	0%	−2%
EMO	0%	0%	−3%	−1%	−1%	1%	−2%	0%	−1%	0%	0%
SENS	−14%	5%	−15%	−6%	−12%	1%	−1%	1%	−6%	0%	0%
FILT	−5%	−2%	−2%	−8%	0%	0%	1%	1%	1%	0%	−1%

To give some indication of the impact of the estimation strategy, table 7 shows observed and estimated probabilities per characteristic over all surveys. Some of the probabilities were lowered, like difficult language in question (DLQ), and others were lifted, like sensitive information (SENS).

In this study, we cannot compare to a complete case analysis. To get some sense of the robustness of the estimation strategy, we omitted five blocks at random for the coders that worked on most of the surveys, shown in the light gray blocks in table 3. Table 8 shows the change in estimated survey probabilities $\hat{p}_s$ relative to all coder data in table 5. For the surveys where a block was omitted, estimates sometimes changed considerably, especially for difficult language in the question (DLQ) and sensitive information (SENS). Five estimates lie outside the original 95% confidence intervals. Hence, estimates are sensitive to the coder allocation scheme, and it is advisable that workload is more evenly spread than was done in our study.

For each of the 2,470 items, an individual estimate $\hat{p}_{s,i}$ is available per item characteristic. These are not shown of course, but in future research, they will be added as explanatory variables in multi-level models to investigate answering behavior on single items, e.g., social desirable answering, acquiescence, underreporting, and do-not-know answers.

4. DISCUSSION

We present methodological tools for a relatively inexpensive and fast preliminary assessment of measurement error risk in surveys. Application of these tools may trigger and inform in-depth cognitive testing and/or experimentation in early (re)design stages. All items of a series of surveys are coded on characteristics that are assumed to be relevant to measurement error. Each survey is assumed to be handled by a limited number of coders. By estimating missing coding data for all other coders for each survey, questionnaire profiles can be constructed. Although the coding is less extensive than testing and experimentation, it does imply an investment in time and costs. The tools must, therefore, be viewed as an addition to the existing toolbox in total quality control. Furthermore, the profiles may be restricted to a subset of the questionnaire modules when survey items have different importance.

In our case study, we focused on six item characteristics that are selected for their relevance to mode-specific measurement errors: difficult language in question, difficult language in answer, risk of non-centrality, sensitive to emotions, sensitive information, and presumed filter question. For each characteristic, the questionnaire profiles for eleven surveys showed the percentage of all items for which the characteristic would be present. For instance, the questionnaire profile for the Labor Force Survey showed that the characteristics difficult language in question and presumed filter question appeared to be present for relatively many items, according to multiple coders, while the other characteristics did not. This implies that there may be a measurement risk coming from inability of respondents to answer questions or motivated underreporting, a risk that in practice may be mediated by the assistance of interviewers. When the questionnaire profile of Politics and Values survey is inspected, it follows that the characteristics sensitive to emotions and sensitive information are relatively present. These characteristics point at measurement risk due to socially desirable answering, which may be stronger in the presence of interviewers. In sum, the questionnaire profiles can be used as a starting point for in-depth cognitive testing and experimental studies.

When using questionnaire profiles specifically as a basis to further investigate measurement error, however, caution is urged for two reasons. First, we observed a large variability in the assigned codes between coders. Some coders were generally conservative in coding a characteristic as present, while other coders were generally liberal in doing so. This variability results in large within standard errors accompanying the estimated probabilities and points to a large uncertainty in the judgment of the presence of the characteristics. Second, the estimated probabilities and profiles may be biased due to a selective clustering of items within surveys. Coders may only work on the items of a restricted amount of surveys, and, in the case of selective clustering, may have coded only a small number of items with certain characteristics. Our case study clearly had a suboptimal design in that two coders only coded the items of two surveys, and three coders coded the items of only one survey. In general, however, our imputation method is a useful extension of the method with only actual coding data, giving a more informative estimation of a questionnaire profile.

Apart from the methodology, the application in this paper may be relevant to questionnaire designers; the surveys included in the study are general purpose surveys with topics that are used in many countries and in many settings (cross-sectional or panel). Other questionnaire designers may perform similar exercises and compare their profiles to ours. We plan to extend the list of surveys and to investigate how to best assign surveys to coders.

The methodology in this paper may be extended in various ways. Future research may focus on optimal coder allocation schemes and ways to make trade-offs between coding hours and coding accuracy. The relatively low agreement between coders and the resulting uncertainty in item characteristic probabilities are reasons for concern and further research. Given a specified accuracy of questionnaire profiles, the number of coders must be larger when agreement is smaller. Three of the coders in our case study are questionnaire experts, while the other coders had no or only some experience with questionnaire design. The agreement between coders did not vary between experts and non-experts, but the expert coders had a bigger vote in choosing and defining the characteristics. Future research may attempt to improve agreement while maintaining relevance in the definition of the item characteristics. In the case study of this paper, standard errors were relatively large, so that normal approximation was invalid and confidence intervals had to be cut off at zero. More efficient coding schemes may remove the need to have better approximations. Nonetheless, future research may address this methodological issue. Finally, future research may also extend profiles to reflect the order in which survey items are posed, so that survey designers may assess the risk of context and order effects. Context effects may be viewed as the impact of the characteristics of all preceding items on answering behavior. Currently, profiles are independent of the order of items.

Appendix A: Questionnaire Profiles

We include the questionnaire profile estimates for the surveys not shown in section 3.2. In all figures, the item characteristics are organized from left to right and from top to bottom: difficult language question, difficult language answer, centrality, sensitive to emotions, sensitive information, and presumed filter question. The light gray boxes show 95 confidence intervals and are based on a normal approximation.

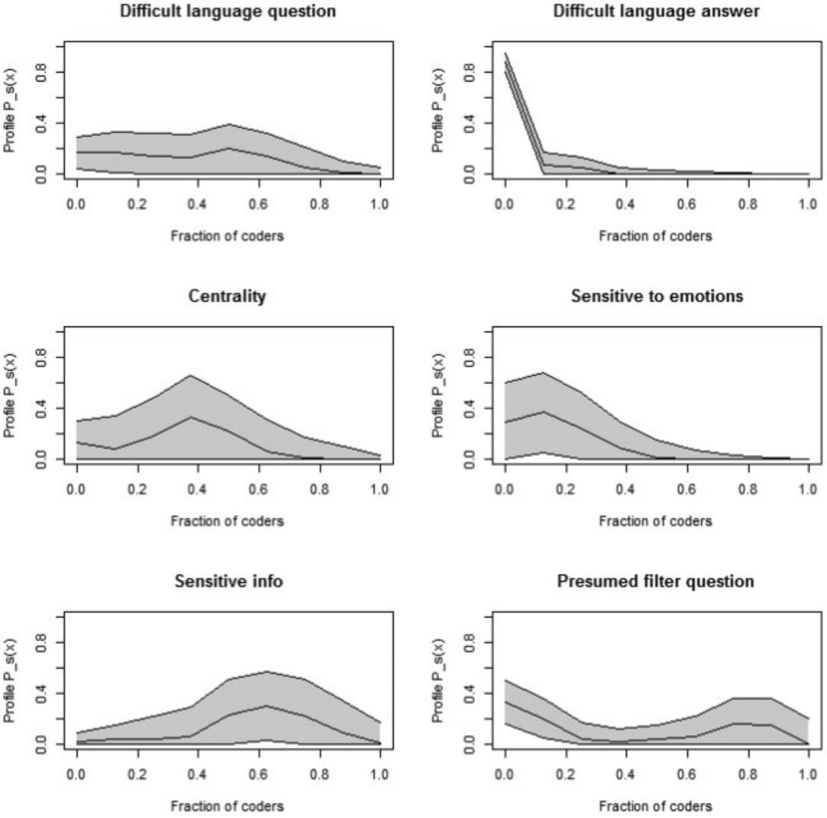

Figure A.1 Questionnaire Profile for Core study Assets (AS).

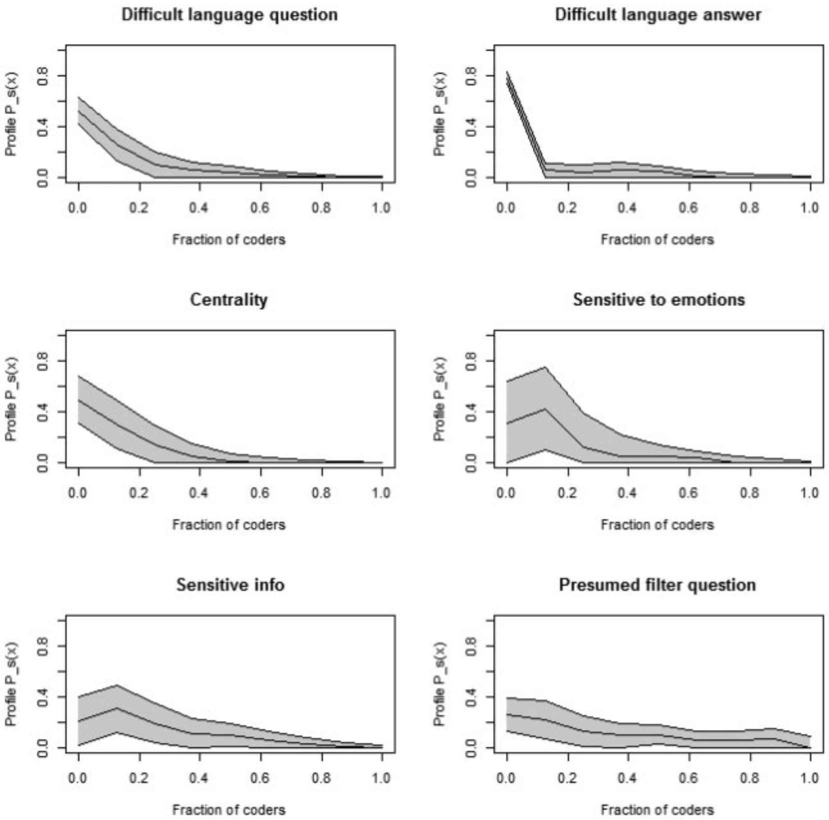

Figure A.2 Questionnaire Profile for Core Study Family and Household (FA).

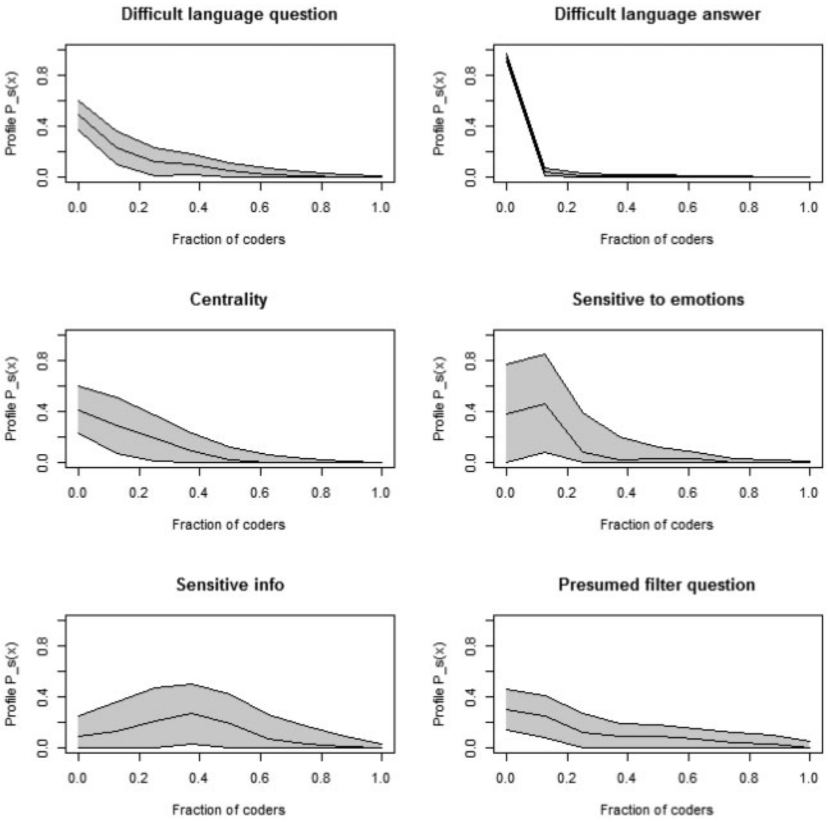

Figure A.3 Questionnaire Profile for Core Study Health (HE).

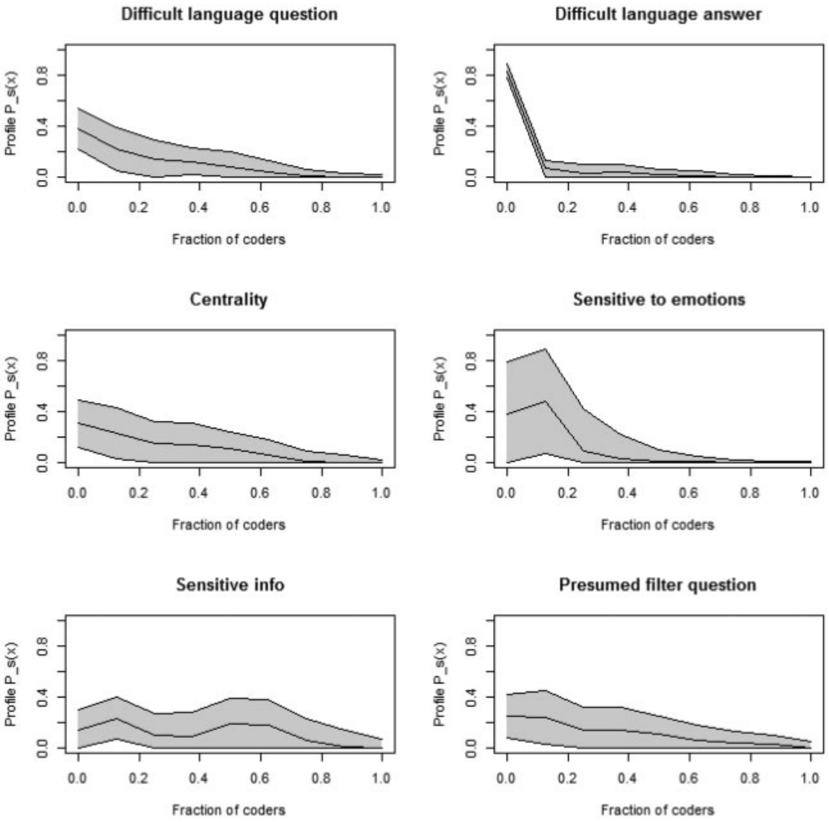

Figure A.4 Questionnaire Profile for Core Study Housing (HO).

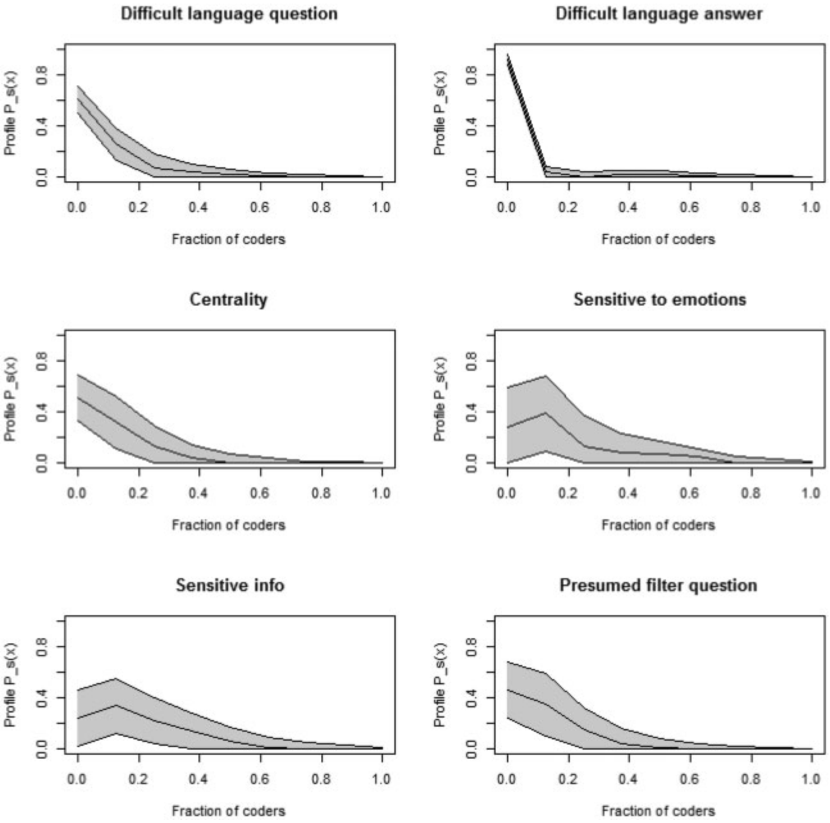

Figure A.5 Questionnaire Profile for Core Study Personality (PE).

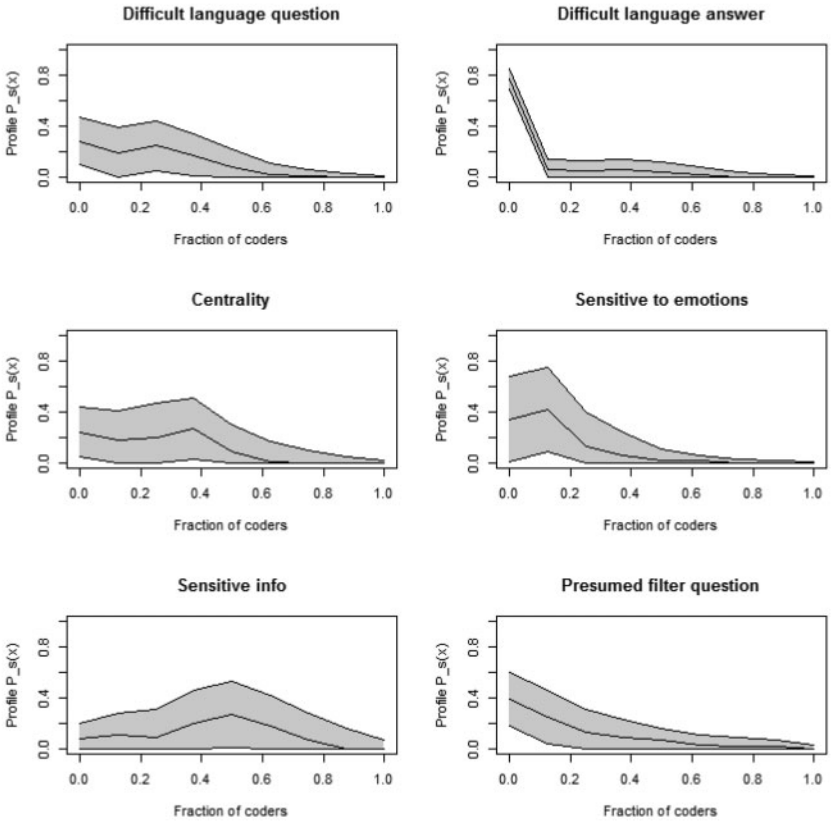

Figure A.6 Questionnaire Profile for Core Study Religion and Ethnicity (RE).

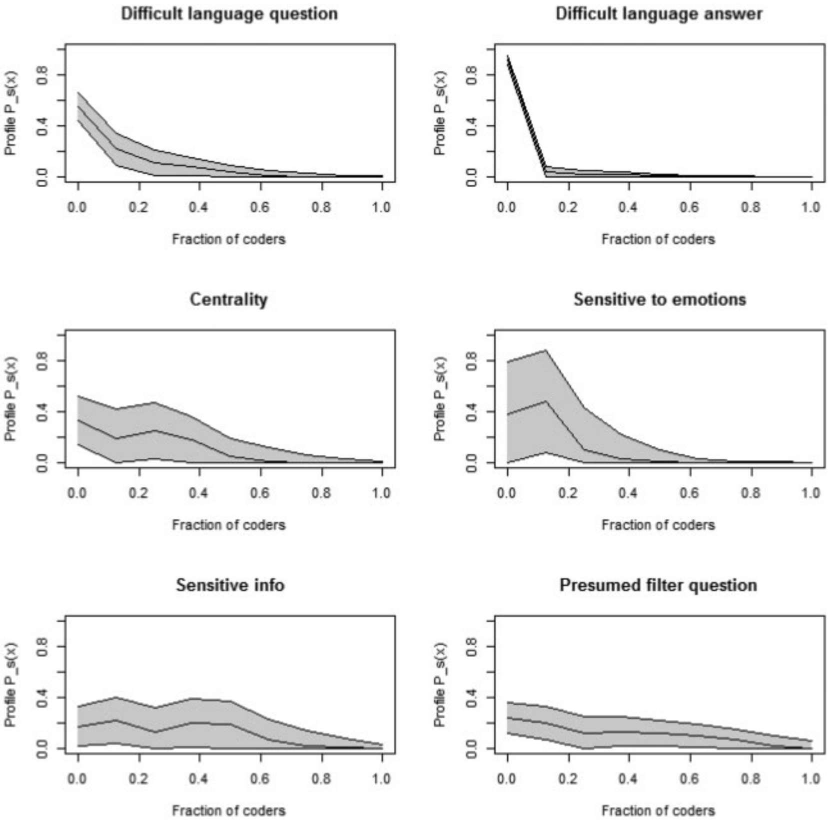

Figure A.7 Questionnaire Profile for Core Study Social Integration and Leisure (SO).

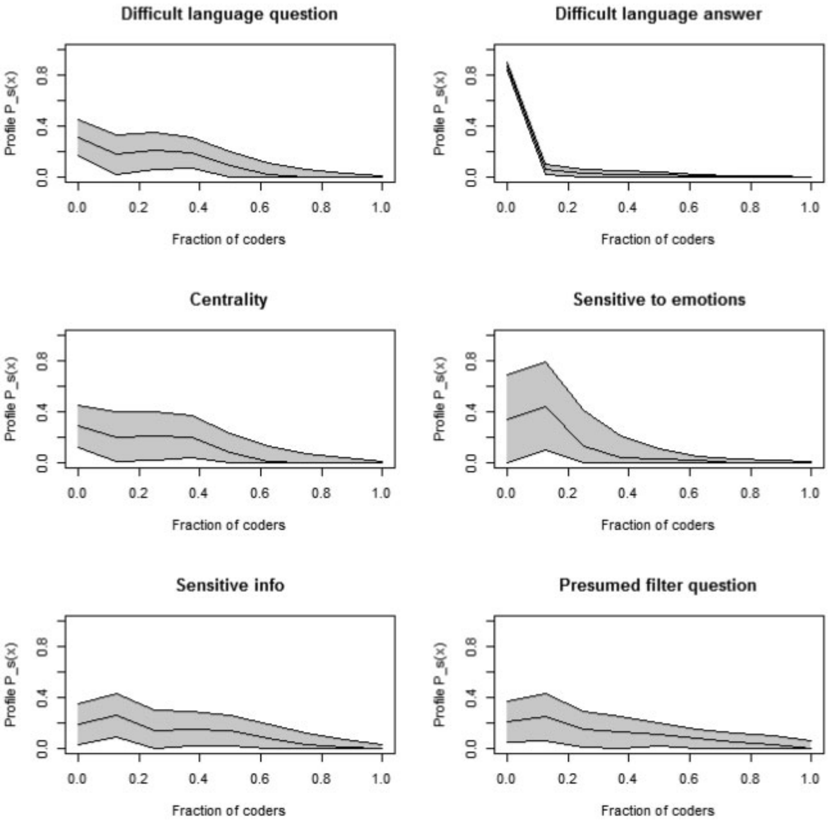

Figure A.8 Questionnaire Profile for Core Study Work and Schooling (WO).

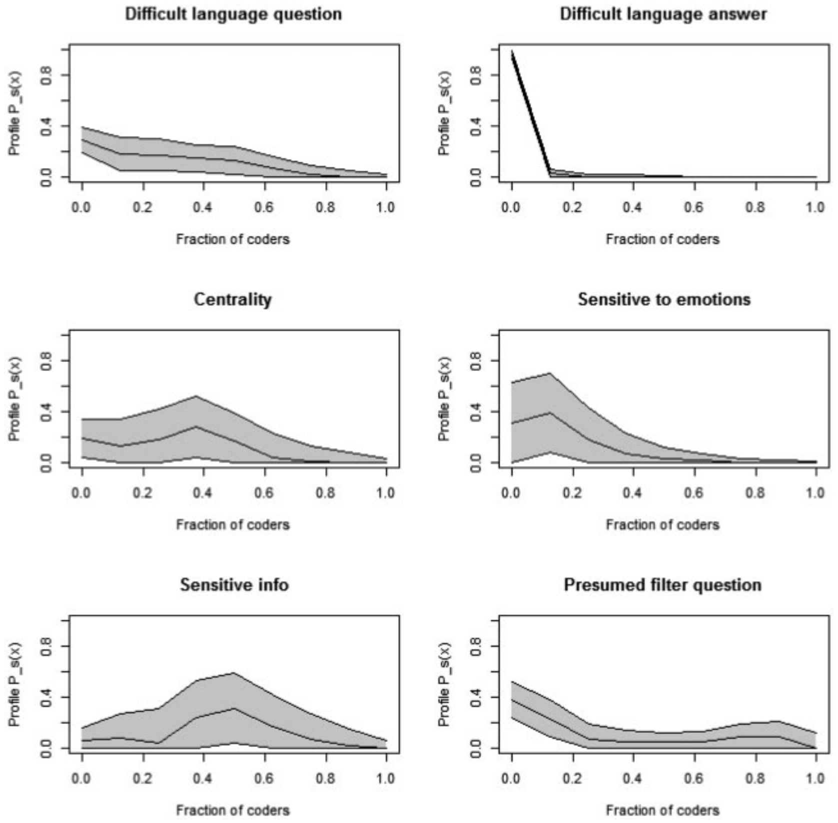

Figure A.9 Questionnaire Profile for Core Study Income (IN).

References

Alwin, D. F., and J. A. Krosnick (1991), "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes," *Sociological Methods and Research*, 20, 138–181.

Bais, F., B. Schouten, P. Lugtig, V. Toepoel, J. Arends-Toth, S. Douhou, N. Kieruj, M. Morren, and C. Vis (2015), "Can Survey Item Characteristics Relevant to Mode-Specific Measurement Error be coded Reliably?" Discussion Paper 201522, Statistics Netherlands, available at www.cbs.nl.

Beukenhorst, D., B. Buelens, F. Engelen, J. Van der Laan, V. Meertens, and B. Schouten (2013), "The Impact of Survey Item Characteristics on Mode-Specific Measurement Bias in the Crime Victimisation Survey," Discussion Paper 201416. Statistics Netherlands, The Hague, available at www.cbs.nl.

Biemer, P. P., R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (1991), *Measurement Error in Surveys*, Hoboken, New Jersey: John Wiley & Sons.

Campanelli, P., G. Nicolaas, A. Jäckle, P. Lynn, S. Hope, M. Blake, and M. Gray (2011), "A Classification of Question Characteristics relevant to Measurement (Error) and Consequently

Important for Mixed Mode Questionnaire Design," paper presented at the Royal Statistical Society, October 11, London, UK.

Cohen, J. (1960), "A Coefficient for Agreement for Nominal Scales," *Education and Psychological Measurement*, 20, 37–46.

Couper, M. P., and F. Kreuter (2013), "Using Paradata to Explore Item-Level Response Times in Surveys," *Journal of the Royal Statistical Society Series A*, 176, 271–286.

Eckman, S., F. Kreuter, A. Kirchner, A. Jäckle, R. Tourangeau, and S. Presser (2014), "Assessing the Mechanisms of Misreporting to Filter Questions in Surveys," *Public Opinion Quarterly*, 78, 721– 733.

Fleiss, J. L. (1971), "Measuring Nominal Scale Agreement among Many Raters," *Psychological Bulletin*, 76, 378–382.

Fowler, F. J. Jr (1995), "Improving Survey Questions: Design and Evaluation," *Applied Social Research Methods Series*, vol. 38. Thousand Oaks, CA: Sage Publications.

Graesser, A. C., Z. Cai, M. M. Louwerse, and F. Daniel (2006), "Question Understanding Aid (QUAID): A Web Facility that Helps Survey Methodologists Improve the Comprehensibility of Questions," *Public Opinion Quarterly*, 70, 1–20.

Kreuter, F., S. Presser, and R. Tourangeau (2008), "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity," *Public Opinion Quarterly*, 72, 847–865.

Olson, K. (2010), "An Examination of Questionnaire Evaluation by Expert Reviews," *Field Methods*, 22, 295–318.

Olson, K., and J. D. Smyth (2015), "The Effect of CATI Questions, Respondents and Interviewers on Response Time," *Journal of Survey Statistics and Methodology*, 3, 361–396.

Presser, S., and J. Blair (1994), "Survey Pretesting: Do Different Methods Produce Different Results?," *Sociological Methodology*, 24, 73–104.

Saris, W. E., and I. Gallhofer (2007), "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions," *Survey Research Methods*, 1, 29–43.

Schouten, B., J. V D. Brakel, B. Buelens, J. V D. Laan, and L. T. Klausch (2013), "Disentangling Mode-Specific Selection and Measurement Bias in Social Surveys," *Social Science Research*, 42, 1555–1570.

Shoukri, M. M. (2010), "Measures of Interobserver Agreement and Reliability," on *CRC Biostatistics Series* (2nd ed.). Boca Raton, TX: Chapman and Hall.

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

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

Van der Vaart, W., J. Van der Zouwen, and W. Dijkstra (1995), "Retrospective Questions: Data Quality, Task Difficulty, and the Use of a Checklist," *Quality and Quantity*, 29, 299–315.

Yan, T., and R. Tourangeau (2008), "Fast Times and Easy Questions: The Effects of Age, Experience, Question Complexity on Web Survey Response Times," *Applied Cognitive Psychology*, 22, 51–68.