# Predicting measurement error variance in social surveys

Daniel L. Oberski
Department of Social Sciences Methodology & Statistics, Utrecht University and Department of Methodology and Statistics, Tilburg University
d.l.oberski@uu.nl

Anna DeCastellarnau
Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra and Department of Methodology and Statistics, Tilburg University and Sociometric Research Foundation
annadecastellarnau@gmail.com

**RECSM Working Paper Number 63**
**September 2021**

# Predicting measurement error variance in social surveys

**Abstract**

Social science commonly studies relationships among variables by employing survey questions. Answers to these questions will contain some degree of measurement error, distorting the relationships of interest. Such distortions can be removed by standard statistical methods, when these are provided knowledge of a question's measurement error variance. However, acquiring this information routinely necessitates additional experimentation, which is infeasible in practice. We use three decades' worth of survey experiments combined with machine learning methods to show that survey measurement error variance can be predicted from the way a question was asked. By predicting experimentally obtained estimates of survey measurement error variance from question characteristics, we enable researchers to obtain estimates of the extent of measurement error in a survey question without requiring additional data collection. Our results suggest only some commonly accepted best practices in survey design have a noticeable impact on study quality, and that predicting measurement error variance is a useful approach to removing this impact in future social surveys.

**Keywords**

machine learning; measurement error; prediction, multitrait-multimethod; SQP

**Acknowledgments**

The authors thank Willem Saris for support and comments; the American Association for Public Opinion Research (AAPOR) for its support of a previous version of this work; and Wiebke Weber, Melanie Revilla and Diana Zavala-Rojas for comments on an earlier version of this manuscript.

## 1. Introduction

Social and behavioral researchers seek to explain opinions, attitudes, and behavior. To achieve this aim, even in the age of big data, the survey question remains the measurement instrument of choice. After all, humans are unique among research subjects in their ability to describe their thoughts and experiences, and a straightforward and cost-effective way to understand them is to ask questions. The answers to these questions will, inevitably, contain measurement error, as survey researchers are well aware (Alwin 2007; Groves 1989; Groves et al. 2009; Saris and Gallhofer 2007; Tourangeau, Rips, and Rasinksi 2000): a serious problem, because measurement error is well-known to bias estimates of relationships between variables of primary interest (Fuller 1987; Gustafson and Greenland 2014; Loken and Gelman 2017).

However, not all survey questions are created equal. When designing a question, researchers determine, often implicitly, many important characteristics. For instance, the language, topic, presence of an interviewer, wording, length, and answer scale are all thought to impact the way questions are answered (Krosnick 1999; Krosnick and Presser 2010). In spite of a range of advice regarding good practices (Dillman, Tortora, and Bowker 1998; Revilla, Zavala-Rojas, and Saris 2016; Saris and Gallhofer 2014; Schuman and Presser 1981; Sudman and Bradburn 1983), generalized knowledge across a range of study topics, countries, and languages is lacking, hindering social science. Furthermore, statistical adjustment for the effects of measurement error requires knowledge of the size of error variances (Carroll et al. 2006; DeCastellarnau and Saris 2014; Fuller 1987; Lilleoja and Saris 2015). Currently most studies lack this information (Saris and Revilla 2016).

The goal of social and behavioral research is, thus, threatened by measurement error in survey data. Estimating the extent of such error is therefore of primary importance, for two reasons: (1) before data collection, to learn how to design better survey questions to minimize the extent of errors, and (2) after data collection, to statistically correct the results for the impact of measurement error on estimates of relationships between variables.

Previous work on estimating the extent of measurement error in survey questions goes back several decades (Groves 2004). In psychology, Lord and Novick (1968) introduced the classical test and item response theory for the estimation of measurement error variance from internal consistency designs (Cronbach 1951; Greene and Carmines 1980). Fellegi (1964) and Hansen, Hurwitz, and Bershad (1961) discussed the estimation of random survey error variance from test-retest designs, while Heise (1969), Jöreskog (1970) and Wiley and Wiley (1970) put such models into a structural equation modelling (SEM) framework, and extended them to a multiple occasions design into the quasi-simplex model (see also: Alwin (2007)). At the same time, Andrews (1984) and Goldberger and Duncan (1973) applied SEM to the multitrait-multimethod (MTMM) design (Campbell and Fiske 1959) to estimate random and correlated survey measurement error variance. Saris and Andrews (1991) extended this SEM model to the true-score multitrait-multimethod (TS-MTMM) model which allows estimating random, systematic and correlated measurement error (Saris and Gallhofer 2014). For categorical data, methods for the estimation of classification error were discussed by Hagenaars (1993) and Lazarsfeld (1950), see also: Biemer (2011). Finally, when gold standard validation data are available, measurement error can be observed directly as the difference between survey answers and true records (Kreuter, Mueller, and Trappmann 2010; West and Little 2013).

These approaches are general methods of estimating measurement error variance in specific survey questions from data collected using a specific design. However, they all share some disadvantages: First, each approach requires supplementary data collection beyond the single survey question under evaluation, as well as additional data analysis. Second, all approaches are post-hoc, meaning that, when measurement error turns out to be high, it is already too late to improve the survey question. Third, validation data is often difficult to obtain, and, once procured, may even not turn out to be error-free (Bakker 2012; Groen 2012; Oberski et al. 2017). Finally, if the impact of measurement error on analyses across the social and behavioral sciences is to be removed, estimates of survey measurement error variance must generalize. To date, however, this generalizability has not been evaluated – leading to limited application of methodological findings to the improvement of social and behavioral research.

We overcome these disadvantages by predicting the extent of measurement error variance in survey questions using random forests of regression trees and validating the results across survey topics and countries. To do so, we employ a meta-dataset of multiple experimental studies created over three decades. This dataset provides, on the one hand, estimates of the extent of measurement error for thousands of questions included in MTMM experiments, and, on the other hand, provides design characteristics of each question. In addition, to improve the survey-based research that social sciences commonly rely on, this paper is accompanied by open source software ("Survey Quality Predictor"; SQP) that allows researchers to predict the extent of measurement errors. Our results before data collection, (1) give validated and quantified measures of the importance of survey questions' design characteristics and (2) allows improving survey questions by comparing the quality obtained with different design characteristics. After data collection they (3) allow correction of statistical results for the biasing impact of measurement error.

Previous studies relating question characteristics to question measurement error variance were performed by Alwin (1997, 2007), Andrews (1984), Költringer (1995), Revilla and Ochoa (2015), Revilla, Saris, and Krosnick (2014), Rodgers, Andrews, and Herzog (1992), Saris and Gallhofer (2014), Saris et al. (2010) and Scherpenzeel and Saris (1997). The primary focus of these studies has been on *parameter estimation*:

generalities regarding the connection between a question's design and its measurement error extent were sought in the interest of furthering survey methodology. Unfortunately, such generalities have proven hard to discover because the set of survey questions considered in each study was limited in topic and country of administration.

The current study incorporates the experiments analyzed by some those authors, as well as a large set of experiments performed in the European Social Survey (ESS; Jowell et al. (2007)). Putting together these studies, a larger and richer data set is available, including an unprecedented number of questions (3,483 in total) and countries (22 in total; Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Great Britain, Greece, Ireland, Netherlands, Norway, Poland, Portugal, Spain, Slovakia, Slovenia, Sweden, Switzerland, Ukraine and United States). Moreover, this study is the first to collect and integrate this large amount of information about commonly used survey questions of the social sciences and includes many more question characteristics than had been used by previous studies. Furthermore, our study differs from previous work in that it focuses on *prediction* rather than parameter estimation, to provide a free and open-source tool, the SQP, that sets of results of direct relevance to researchers in the social and behavioral science.

In section 2, we first present the type of survey questions, measurement error indicators, and question characteristics included in this study. In section 3, we describe the prediction model. Next, the results obtained are summarized in section 4, and its application to survey research is illustrated in section 5. Finally, section 6 concludes.

## 2. Data

Our meta-dataset comprises results from 98 survey experiments administered to a total of 174,245 respondents (about 1,300 on average per experiment), during the period 1979-2006 across 22 countries. In total, the dataset consists of 3,483 survey questions. For each survey question, we obtained:

1. estimates of two indicators of measurement error from TS-MTMM models; and

2. codes of the question characteristics describing the choices made by the survey researcher when designing each question (design characteristics).

In this section, we briefly describe the selection of survey questions into our study, as well as the methods used to obtain question measurement error indicators and characteristics. More in-depth descriptions of both can be found in: (Jowell et al. 2007; Költringer 1995; Saris et al. 2011; Saris and Gallhofer 2014; Scherpenzeel and Saris 1997)

### 2.1 Survey question selection

The survey questions included in our experiments were administered to random probability samples of respondents as part of existing academic social science studies (full documentation for the largest of these studies, the ESS, is available at http://www.europeansocialsurvey.org/data). A wide range of topics was studied, including political, economic and work issues, subjective evaluations, crime, values, behavioral activities and feelings. Table A1, in the appendix, contains further details on these topics, and other background characteristics of the experiments, including the country, year, mode of data collection, experimental design of the studies, the number of questions evaluated, and the sample size.

### 2.2 Estimation of measurement error indicators

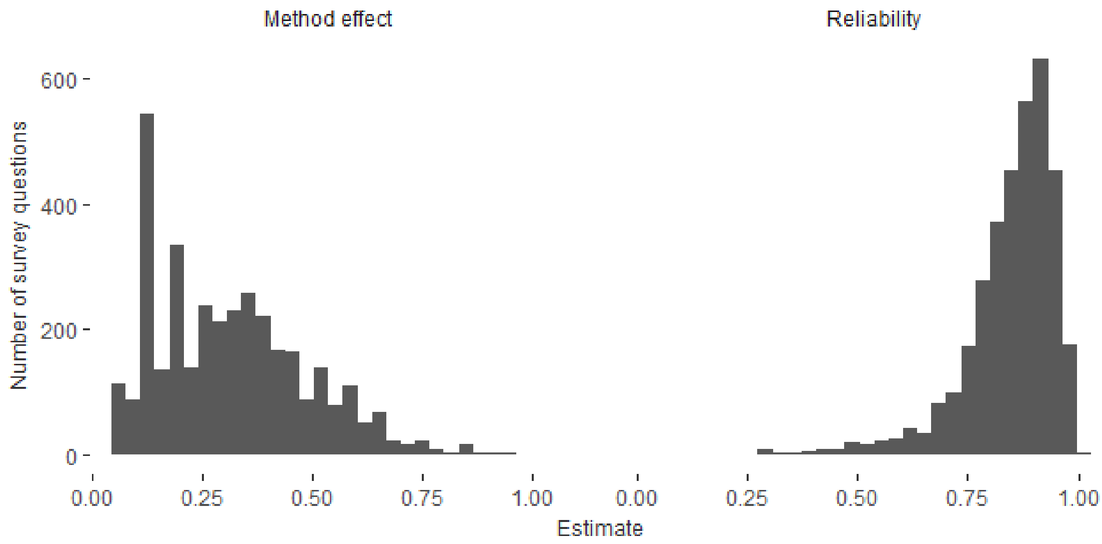
An abundance of survey measurement error indicators exists (Biemer and Lyberg 2003; Madans et al. 2011). Because social and behavioral research is often correlational, in this study we focus on quantifying how strongly such correlations are biased due to potential errors in the survey. This corresponds to the amount of random measurement error variance, biasing correlations downwards, and correlated measurement error, biasing them upwards. Note, however, that in multi-variable settings, such as linear regression, the direction of the bias is not predetermined: random error may bias regression coefficients downwards or upwards, for instance (Loken and Gelman 2017). Quantifying random and correlated measurement error is therefore an essential precondition to valid social and behavior science.

The TS-MTMM approach (Saris and Andrews 1991) was designed to disentangle these two sources of bias using a classical within-persons design due to Campbell and Fiske (1959). In this design, as applied to survey research (Alwin and Krosnick 1991; Andrews 1984), respondents are asked the same question measuring correlated concepts of interest (*traits*) repeatedly, using different design characteristics (*methods*). For example, the question “How satisfied are you with your job?”, might first be asked using a scale from 0, “Not at all”, to 10, ‘Completely”, and subsequently – after a sufficient interval (van Meurs and Saris 1990) – with a scale from 1, “Completely dissatisfied” to 5, “Completely satisfied”. In this example, the scale types used are the *multiple methods* in the MTMM. In principle, one repetition of the same question is required per method; in practice, however, each respondent is burdened with a single repetition only (Saris, Satorra, and Coenders 2004). Using these same methods to ask further questions, for instance about satisfaction with health and relationship, leads to the *multiple traits*.

The MTMM data collection scheme enables statistical modeling using the TS-MTMM model (Saris and Andrews 1991). This model regards each question as a cross between a trait and a method and estimates random trait, method, and error effects. Thus, the TS-MTMM model disentangles and quantifies each survey question's proportion of random and correlated measurement error variance. The complement of the proportion of random error variance is commonly known as *true-score reliability*. An alternative interpretation of this proportion is that it is the theoretical correlation between the true value and the observed answer; that is, the multiplicative downwards bias in correlation among observed variables (Fuller 1987). The complement of the proportion of systematic error (method) variance is known in the MTMM literature as *true-score validity*. An alternative interpretation of this proportion is that it is the hypothetical spurious correlation between two otherwise unrelated questions asked in the same way, also known as *method effect*; that is, it is the upward bias in a correlation due to the method (Saris and Gallhofer 2014). To avoid confusion with other common uses of the term *validity*, we will adopt the terms *reliability* and *method effect* here. For an extensive treatment of the TS-MTMM approach to measurement error evaluation, we refer to: Saris and Gallhofer (2014).

We obtained data from 98 MTMM experiments performed over the period between 1979 and 2006. TS-MTMM structural equation models (SEM; Bollen (1989)) were estimated using the LISREL software (Jöreskog and Sörbom 1996) to obtain estimates of the reliability and method effect of the 3,483 survey questions experimented upon. Figure 1 shows the distribution of these estimates in our data. Reliability was generally high (close to unity). Method effect was generally low. Although most survey questions exhibited a reasonable quality in terms of these indicators, there was considerable variability. This variability may be partly explained by differences in the way the questions were asked, i.e. in their characteristics. Moreover, the distribution of method effect estimates was highly skewed, whereas reliability showed long tails to the left with a small number of very unreliable questions.

**Figure 1. Distribution of the method effect and reliability estimates (in the original scale) obtained for n = 3,483 survey questions**




To facilitate explanation of these highly non-normal variables, we transformed them using the logit transform, $x' = \log(\frac{x}{1-x})$, where $x$ refers to either the reliability or the method effect (both between 0 and 1) and $x'$ to its logit transform (a real number). In our prediction models, we predict these transformed quantities and back-transform them after prediction to obtain predictions of the original quantities.

### 2.3 Characteristics of survey questions

The diversity of survey question characteristics affecting measurement error has been suggested in the rich literature on survey methodology (Alwin 2007; Alwin and Krosnick 1991; Belson 1986; Krosnick and Fabrigar 1997; Rodgers, Andrews, and

Herzog 1992; Saris and Gallhofer 2014; Schuman and Presser 1981). For example, questions with fewer scale points give the respondent less freedom of expression, potentially reducing reliability (Alwin 1997; Alwin and Krosnick 1991; Krosnick and Presser 2010). However, at the same time, the freedom to express systematic biases may be similarly increased by the use of more scale points, increasing method effect (Alwin 1997; Komorita and Graham 1965; Revilla and Ochoa 2015; Saris and Gallhofer 2014). Other question characteristics of potential importance are the complexity of the language used, the presence of a "don't know" option, the presence or absence of an interviewer, country and language of administration, and so on. A review of many characteristics was given by Saris and Gallhofer (2014). Combining these with several other characteristics a coding system was developed with 73 different codes for each survey question (see coding instructions in: Survey Quality Predictor (2017), for a complete overview of the characteristics used).

The characteristics were coded by a team of trained multilingual coders. A subset of identical questions that were translations of each other were double-coded in English by an expert, the final codes being arrived at by consensus after an adjudication procedure. The full database of survey questions, characteristics (codes), reliability and method effect estimates can be explored by selecting all questions "with MTMM" in the open and free to use web software we provide at: http://sqp.upf.edu/. From this list, of 73 characteristics, some were combined (e.g. the linguistic characteristics) and some other contextual variables (e.g. whether the question was a repetition or whether the question belonged to ESS). With the final meta-dataset, 63 variables were used as predictors in the models. These variables and its distributions are presented in Table A2, in the appendix.

## 3. Method

Our goal is to predict, as accurately as possible, the reliability and method effect of survey questions from the characteristics of those questions included in the meta-dataset. In selecting an appropriate predictive model, two issues with the data must be accounted for. First, the characteristics are highly correlated. For instance, questions with fewer scale points are more likely to fully label those points, and showcards are more commonly used in face-to-face than in other data collection modes. Second, a host of possible interaction effects (nonlinear prediction curves) may be expected. For example, “don’t know” and neutral categories are known to interact in their impact in quality (Sturgis, Roberts, and Smith 2014) and providing respondents with a concise explanatory introduction may be helpful, whereas providing an overwhelming amount of information may increase measurement error. To account for characteristic multicollinearity and possible interactions, we therefore selected random forests of regression trees (Breiman 2001; Hastie, Tibshirani, and Friedman 2009) as an appropriate supervised machine learning strategy.

The random forest algorithm is shown in Box 1. A random forest is an ensemble of binary regression trees (CART; Breiman et al. (1984), constructed by combining bagging (Breiman 1996) with sampling of characteristics that predict the outcome (Breiman 2001). Regression trees allow for the possibly complex interactions between characteristics in predicting reliability and method effect, while bootstrapping combined with predictor sampling prevents overfitting and multicollinearity by decorrelating the trees over bootstrap samples. The final prediction is an aggregation of the ensemble of trees.

**Box 1. Pseudo-code for the random forest algorithm applied to measurement error prediction**

**Growing** a random forest:

WHILE number of trees grown < *ntrees*, DO:

1. From the 3,483 questions, select 3,483 questions at random with replacement to create the "in-bag" dataset. Those questions not selected are considered "out-of-bag";
2. Select a number of characteristics at random (*mtries*) from the 64 question characteristics of the in-bag dataset to predict the logit-transformed reliability/method effect;
3. Grow a regression tree of at most deep (*max_depth*) from the in-bag dataset using the *mtries* selected characteristics;
4. In out-of-bag and validation datasets, predict the logit-transformed reliability/method effect from the tree grown and calculate validation metrics.

RETURN the ensemble containing *ntrees* bagged CART trees.

**Predicting** logit-transformed reliability/method effect of a question from a random forest:

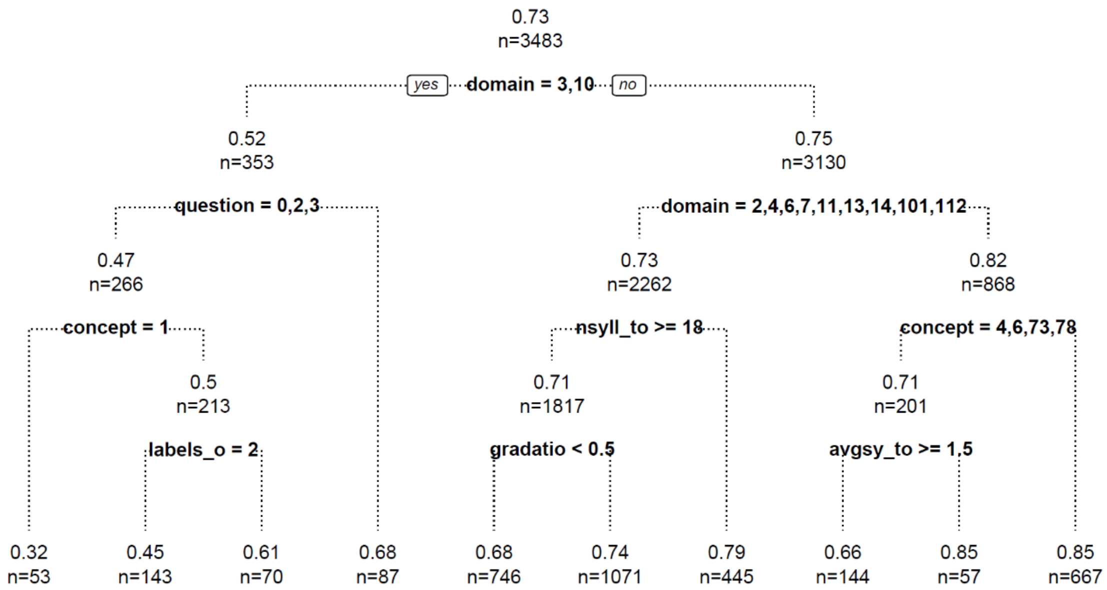
FOR each tree in the ensemble:

Predict the logit-transformed reliability/method effect for the question using the tree.

RETURN the average of all *ntrees* predictions.

For example, suppose we have the survey item "Please tell me to what extent your lifestyle is beneficial for your health". Using the illustrative example, in Figure 2, of a regression tree for reliability with only 7 characteristics, the prediction starts at 0.73, the overall average reliability of all questions, and differentiates more and more as the leaves are followed downwards. A question such as this, which asks about 'Health' (domain = 3), is not interrogative (questiontype ≠ 1), and asks about an "Evaluative belief" (concept = 1) is predicted to have a low reliability of 0.32 (leftmost node). This prediction would be based on 53 observations of our dataset. However, a much higher reliability prediction, of 0.85 (rightmost node), would be obtained if the question asked about a "Feeling" (concept = 2) about the "National institutions" (domain = 103), such as “are you satisfied with the way Parliament is doing its job?”. The final prediction obtained for this question is an average of the total number of trees used which contain different number of sampled characteristics.

**Figure 2. Illustrative example of a single regression tree predicting reliability form the question characteristics**




As indicated in Box 1, several tuning parameters must be selected: the number of characteristics sampled (*mtries*), the number of trees in the ensemble (*ntrees*), and the maximal depth of each tree (*max_depth*). We performed a grid search to select these by comparing the mean square prediction error in a validation sample over all combinations of a range of possible options. Based on these results we grew 1,500 regression trees, sampling 20 characteristics for each tree with a maximum depth of 30, and 20 characteristics with maximum tree depth 15 to predict the logit-transformed reliability and method effect, respectively.

To predict the two dependent logit-transformed variables, reliability and method effect, we estimated two separate prediction models, yielding separate random forests.

The data were split into a 60% training sample, 20% validation sample, and 20% testing sample. The algorithms were run using the packages randomForest version 4.6-12 (Liaw and Wiener 2002) and h2o version 3.10.3.6 (Aiello, Kraljevic, and Maj 2015) in R version 3.3.3 (R Core Team 2017).

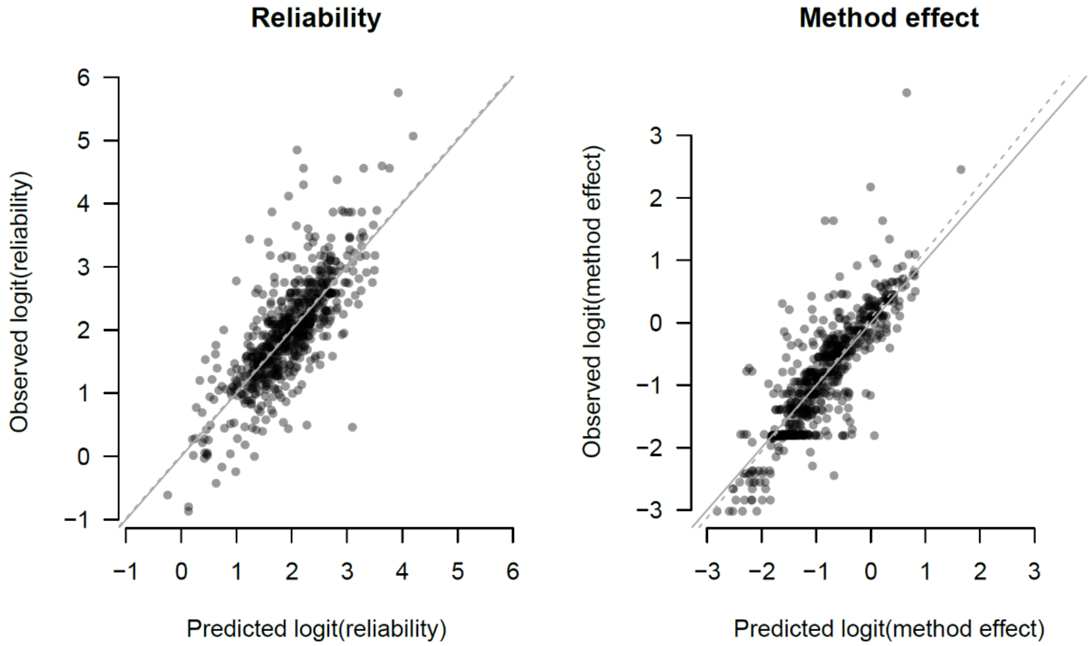
## 4. Results

### 4.1 Prediction quality

After running the random forest algorithm on the training set, we obtain validation metrics based on the test set – the 20% of the "out-of-bag" questions not selected to train the models. These indicate that the reliability and method effect, of new questions, could be predicted relatively well on data held out from the prediction model, with squared correlations ($R^2$) between predicted and observed logit-transformed coefficients out-of-sample of $R^2 = 0.60$ for reliability, and $R^2 = 0.71$ for method effect. The out-of-sample median absolute errors (*MAE*) on the original (0, 1) scales were *MAE* = 0.03 for reliability and *MAE* = 0.04 for method effect. This prediction can be considered relatively accurate.

To illustrate how well the random forest did at predicting new questions' TS-MTMM reliability and method effect, Figure 3 plots the predictions against the observed values in a 20% test set not used to estimate the model.

**Figure 3. Predicted versus observed logit-transformed reliability and method effect in the test set (n = 697 survey questions)**




As indicated by the $R^2$ values, the observed logit-transformed reliability and method effect estimates cluster around the predictions. The solid line corresponds to exact equality between the two quantities, while the dashed line shows the best fitting regression line. For reliability, the best fitting line sticks to equality, indicating approximately unbiased prediction of reliability. For method effect, the dashed line sticks out below and above the solid line slightly for the extreme questions, indicating slight smoothing of these predictions towards the average.

### 4.2 Question characteristic importance

For the design of social science studies, it will be of interest which characteristics of a question design most impact its measurement properties (DeCastellarnau 2017). Based on the random forest, this question can be answered by calculating permutation importance measures. Permutation importance proceeds by randomizing the values of a single predictive characteristic and recording the drop in mean square prediction error or accuracy due to the loss of information incurred by removing this characteristic from the prediction. This process is then repeated for each predictor, and the predictors can be ordered by their importance based on this measure.

Applying *marginal* permutation importance to our model yields the predictor ordering shown in Table 1, which lists the predictor variables with the ten highest proportional reductions in mean squared error. As shown in this table, the strongest predictors, in this sense, are highly similar for the reliability and method effect. In particular, the topic of the question (*domain*), the number of categories (*ncategories*), its position in the questionnaire (*position*), the country in which it is asked (*country*), whether it is about feelings, evaluations or other (*concept*), the average number of words, i.e. the length of the question (*avgwrd_total*) and whether the scales' labels are fully or partially labelled (*labels*) are all strongly correlated with both reliability and method effect (see Table A1, in the appendix, for further details on the characteristics). The full table of variable marginal importance results can be found in Table A3, in the appendix.

**Table 1. Ten predictors with highest marginal permutation importance from the full random forest prediction model**

| Rank | Logit Reliability | | Logit Method effect | |
|---|---|---|---|---|
| | Predictor | Propensity MSE | Predictor | Propensity MSE |
| 1 | domain | 0.208 | ncategories | 0.130 |
| 2 | ncategories | 0.165 | position | 0.116 |
| 3 | country | 0.095 | domain | 0.113 |
| 4 | concept | 0.086 | country | 0.070 |
| 5 | position | 0.058 | concept | 0.067 |
| 6 | nsyll_total | 0.035 | labels | 0.045 |
| 7 | avgwrd_total | 0.028 | avgwrd_total | 0.028 |
| 8 | avgsy_total | 0.021 | scale_basic | 0.024 |
| 9 | nwords_total | 0.021 | repetition | 0.022 |
| 10 | labels | 0.017 | scale_corres | 0.021 |

These results show that the contextual characteristics of the experiments, i.e. the *domain*, the *country* and the *concept*, are important when explaining the variations in the measurement quality of survey questions, i.e. a change in the choice of these characteristics can aid to higher change in the measurement quality obtained. However, this kind of characteristics are not varied for the experimental purpose of each experiment, i.e. they are constant. This result can be explained by the structure of the meta-data. As shown in Table A1, in the appendix, more than 70% of the questions belong to the ESS. In the ESS, each MTMM experiment is administrated to more than 15 countries, which means the contextual characteristics of these experiments do not vary within experiment over the 15 to 20 countries.

The variable importance measures in Table 1 are useful to understand how the random forest model predicts the measurement properties of a question. However, they are not useful to potentially understand which question design characteristics a social scientist should consider, for two reasons. First, the importance measures are marginal, i.e. when randomizing the values of some characteristic (A), the predictive value of other characteristics that depends on A is also destroyed, and counted towards A's predictive value (Hothorn, Hornik, and Zeileis 2006). For this reason, Strobl et al. (2007) proposed conditional permutation importance measures that preserve the predictive value of other characteristics. Second, some characteristics that are useful to predict reliability and method effect, and therefore to correct analyses for measurement error, are not under the control of social scientists who design questions: these are the *country*, *domain*, *concept*, future vs. past questions (*future*, *past*), social desirability (*socdesir*), *centrality*, and experimental design characteristics (*repetition*, *from*).

To provide insight into the characteristics of a question that may be important when designing a question, we calculated the conditional importance measures on a random forest of “unbiased conditional inference trees” using the party package version 1.2-2 (Hothorn, Hornik, and Zeileis 2006; Strobl et al. 2007, 2008). In these models, we excluded the above characteristics not under the designer’s control. As expected from the marginal importance measures, the $R^2$ prediction worsened after removing these characteristics relative to the full model to $R^2 = 0.55$ for reliability and $R^2 = 0.70$ for method effect. However, when designing a question rather than predicting its measurement properties, the conditional variable importance measures obtained give useful information on the question characteristics to account for.

Table 2 lists the ten most important question characteristics out of those 54 that can be controlled by the designer. The column “accuracy decrease” lists the absolute mean decrease in accuracy after permuting the corresponding characteristic. The higher overall values for method effect prediction reflect the higher proportion of explained variance ($R^2$) of this model. The full table of variable conditional importance results can be found in Table A4, in the appendix.

**Table 2. Ten predictors with highest conditional permutation importance from the reduced random forest prediction model.**

| Rank | Logit Reliability | | Logit Method effect | |
|---|---|---|---|---|
| | Predictor | Accuracy decrease | Predictor | Accuracy decrease |
| 1 | ncategories | 0.128 | ncategories | 0.170 |
| 2 | scale_corres | 0.052 | labels | 0.121 |
| 3 | labels | 0.051 | position | 0.103 |
| 4 | questiontype | 0.046 | scale_corres | 0.066 |
| 5 | labels_order | 0.032 | range.corres | 0.058 |
| 6 | range.corres | 0.032 | visual | 0.057 |
| 7 | avgwrd_total | 0.031 | scale_basic | 0.052 |
| 8 | stimulus | 0.027 | labels_order | 0.038 |
| 9 | scale_basic | 0.021 | stimulus | 0.037 |
| 10 | fixrefpoints | 0.021 | showc_horiz | 0.034 |

As Table 2 shows, in accordance with the literature (Andrews 1984; Rodgers, Andrews, and Herzog 1992), the number of categories is the most highly predictive of both reliability and method variance. Further question characteristics of high importance to both reliability and method variance are: whether the scale is fully or partially labeled, or lacks labels (*labels*); whether the numbering of the scale points corresponds with their labels (*scale_corres*), whether the semantic ordering of the labels is decreasing or increasing (*labels_orde*r); whether a unipolar response scale has been used for unipolar concepts and a bipolar one for bipolar concepts (*range.corres*); and the position of the question in the questionnaire (*position*).

Research on this has proven the effect of this characteristics on data quality in general. For instance, as summarized in DeCastellarnau (2017), the number of categories has a clear effect on data quality (Alwin 1997, 2007; Alwin and Krosnick 1991; Andrews 1984; Komorita and Graham 1965; Lundmark, Gilljam, and Dahlberg 2016; Preston and Colman 2000; Revilla and Ochoa 2015; Revilla, Saris, and Krosnick 2014; Rodgers, Andrews, and Herzog 1992; Saris and Gallhofer 2014; Scherpenzeel and Saris 1997; Weijters, Cabooter, and Schillewaert 2010), as well as the effect of labelling

(Alwin 2007; Alwin and Krosnick 1991; Andrews 1984; Eutsler and Lang 2015; Krosnick and Berent 1993; Menold et al. 2014; Moors, Kieruj, and Vermunt 2014; Rodgers, Andrews, and Herzog 1992; Saris and Gallhofer 2014; Weijters, Cabooter, and Schillewaert 2010), its' order (Christian, Dillman, and Smyth 2007; Christian, Parsons, and Dillman 2009; Krebs and Hoffmeyer-Zlotnik 2010) or the scales' correspondence (Rammstedt and Krebs 2007; Saris and Gallhofer 2014).

In contrast, there hasn't been much research on the effect of the polarity consistency between the theoretical and the used range of the scale or the scale's position that helps us to prove this result.

Some question characteristics were highly predictive of reliability but are not in the top ten best predictors for method effect. The overall higher explained variance for method effect, however, does not necessarily reflect a worse predictive value of that characteristic. For example, the grammatical form of the question (*questiontype*), which indicates whether the question is interrogative, imperative or declarative, is found to be highly predictive of reliability compared to other characteristics but is not in the top ten for method effect prediction. However, the full list of characteristic accuracy reductions (see Table A4 in the appendix) reveals that permuting this characteristic reduced the accuracy by 0.020 for method effect prediction whereas the corresponding reduction for reliability prediction in Table 2 is 0.046. Thus, if a question designer wishes to optimize reliability, the top ten in Table 2 should be a good place to start, even though relative to the variance in reliabilities found in our data, less can be accomplished in general by such improvements for reliability than for method effect. We hasten to add, however, that, since these question characteristics were largely observational, causal interpretations of their effects must be approached with caution.

Which characteristics are *not* highly predictive of question measurement properties is also of considerable interest to survey methodology and social science more generally. Permuting the precise layout of any showcards, the use of instructions for the respondent or the interviewer, the length of the labels, and the absolute number of abstract nouns, words and subordinate clauses yielded low reductions in prediction accuracy.

## 5. A tool to predict the extent of measurement error in survey questions

Our findings are useful to social and behavioral science in two ways. First, our prediction model can be applied to survey questions used in the field, and the resulting error rate estimates used to remove bias through well-established statistical methods (DeCastellarnau and Saris 2014; Lilleoja and Saris 2015; Saris and Revilla 2016). Second, before researchers go into the field, our predictions of the quality of their intended measurement instruments can be used to improve data collection, thereby improving future studies (Revilla, Zavala-Rojas, and Saris 2016). To facilitate these uses of our study results, we provide a free and open-source tool to the research community ("Survey Quality Predictor"; SQP).

SQP provides the coding system used to generate characteristics of survey questions. Once these codes are given, the software then runs the prediction model to predict reliability and method effect (or its complement: true-score validity), as well as prediction intervals. The system also provides a database of all questions used in this study, their codes, the associated estimates, and can be used to generate new predictions. The SQP source code and its full history is made available in an open repository (see https://github.com/recsm). The SQP is freely accessible online at http://sqp.upf.edu.

## 6. Conclusion and discussion

In spite of a recent surge of interest in using big data from online "digital traces", mobile phones, social media, and administrative registers (Kreuter and Peng 2014; Youyou, Kosinski, and Stillwell 2015), the humble survey question remains a highly effective way to understand human thoughts, attitudes, and behavior. As several authors have noted, combining what people tell us about their lives in surveys with surreptitiously collected big data has enormous potential to advance our understanding of the social world (Japec et al. 2015). When doing so, however, it remains crucial to use the best tools available and to remove the negative impacts of their shortcomings.

This study aimed to contribute to the advancement of social and behavioral science by reducing measurement error and allowing researchers to statistically remove the bias caused by it. Compiling decades of data from survey experiments into a large meta-analytic prediction model, we have shown that survey question reliability and method effects can be accurately predicted from the way the question has been asked. The results of our model can be used to determine better ways to ask future questions as well as correct for the inevitable shortcomings of existing ones. In addition, we provide open-source software that allows the application of our results to future studies that employ surveys.

Our study has several limitations, which remain topics for future work. First, our method of estimating reliability and method effect in survey questions has been the TS-MTMM approach. This approach has as drawbacks that it makes assumptions regarding the time period required to remove memory effects on the estimation, regarding the lack of carry-over effects, and regarding the source of method effects (Alwin 2011). Cernat and Oberski (forthcoming) discuss improvements over the TS-MTMM design that remove all but the first assumption. In our results, however, the fact that a question has been repeated is not a highly important predictor for reliability. Future work would expand on existing research into memory effects (van Meurs and Saris 1990) and investigate whether predictions change substantially as a consequence of incorporating these more modern methods.

Second, codings were human-generated and will therefore contain a certain degree of measurement error themselves, despite efforts to limit such errors through our adjudication procedure and quality control. For this reason, we have focused on prediction rather than interpretation of the estimated parameters. In future, statistical modeling of question reliabilities might additionally incorporate the reliability of the codes themselves.

Third, the question characteristics considered were obtained from literature reviews of existing research into question reliability. Future work might focus on discovering new such characteristics that might relate to question quality – for example using techniques from natural language processing and machine learning such as word embeddings. All preceding issues limit the accuracy of the prediction; thus, the mentioned future improvements can be expected to further improve upon the prediction error of the current study.

Fourth, predictions are based on the data available (see Table A2 in the appendix), which are limited to the type of attitude questions commonly found in social surveys. For example, more novel methods such as anchoring vignettes (Hainmueller, Hangartner, and Yamamoto 2015; King et al. 2004) are not included in our study. Researchers who wish to employ our software for reliability prediction are therefore advised that reliability prediction for questions dissimilar to those used in the public database must rely on extrapolation. Furthermore, prediction intervals can be quite wide, and it is advisable to examine not only the point prediction but also its interval.

Methods are available to account for such uncertainty in measurement error corrections (Bakk, Oberski, and Vermunt 2014; Oberski and Satorra 2013). Similarly, while our study employed experimental data from a large database of social surveys, this database still excludes a substantial group of countries, languages, and topics. Future work should therefore continue the effort to ascertain the quality of social science experimentally.

## References

Aiello, Spencer, Tom Kraljevic, and Petr Maj. 2015. "h2o: R Interface for H2O." http://www.h2o.ai.

Alwin, Duane F. 1997. "Feeling Thermometers versus 7-Point Scales. Which Are Better?" *Sociological Methods and Research* 25(3): 318–40.

———. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, NJ, US: John Wiley and Sons, Inc.

———. 2011. "Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach." In *Question Evaluation Methods: Contributing to the Science of Data Quality*, eds. Jennifer Madans, Kirsten Miller, Aaron Maitland, and Gordon Willis. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Alwin, Duane F., and Jon A. Krosnick. 1991. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods and Research* 20(1): 139–81.

Andrews, Frank M. 1984. "Construct Validity and Error Components of Survey Measures: A Structural Modelling Approach." *Public Opinion Quarterly* 48(2): 409–42.

Bakk, Z., Daniel L. Oberski, and Jeroen K. Vermunt. 2014. "Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference." *Political Analysis* 22(4): 520–40.

Bakker, Bernardus Franciscus. 2012. "Estimating the Validity of Administrative Variables." *Statistica Neerlandica* 66(1): 8–17.

Belson, William A. 1986. *Validity in Survey Research*. Aldershot, England: Grower.

Biemer, Paul P. 2011. *Latent Class Analysis of Survey Errors*. Hoboken, NJ: John

Wiley and Sons, Inc.

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. Wiley.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. John Wiley & Sons.

Breiman, Leo. 1996. “Bagging Predictors.” *Machine learning* 24(2): 123–40.

———. 2001. “Random Forests.” *Machine learning* 45(1): 5–32.

Breiman, Leo, Jerome H. Friedman, R. A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Monterey, CA: Wadworth and Brooks.

Campbell, Donald T., and Donald W. Fiske. 1959. “Convergent and Discriminant Validation by the Multitrait-Multimethod Matrices.” *Psychological Bulletin* 56(2): 81–105.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. Second. Boca Raton, FL: Chapman and Hall/CRC.

Cernat, A., and Daniel L. Oberski. 9999. “Extending the within-Persons Experimental Design: The Multitrait-Multierror (MTME) Approach.” In *Experimental Methods in Survey Research*, ed. P. J. Lavrakas. New York, NY, US: John Wiley & Sons.

Christian, Leah Melani, Don A. Dillman, and Jolene D. Smyth. 2007. “Helping Respondents Get It Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys.” *Public Opinion Quarterly* 71(1): 113–25.

Christian, Leah Melani, Nicholas L. Parsons, and Don A. Dillman. 2009. “Designing Scalar Questions for Web Surveys.” *Sociological Methods & Research* 37(3): 393–425.

Cronbach, Lee J. 1951. “Coefficient Alpha and the Internal Structure of Tests.” *Psychometrika* 16(3): 297–334.

DeCastellarnau, Anna. 2017. “A Classification of Response Scale Characteristics That Affect Data Quality: A Literature Review.” *Quality & Quantity* Online: 1–37. http://link.springer.com/10.1007/s11135-017-0533-4.

DeCastellarnau, Anna, and Willem E. Saris. 2014. “A Simple Way to Correct for Measurement Errors in Survey Research.” *European Social Survey Education Net (ESS Edunet)*. http://essedunet.nsd.uib.no/cms/topics/measurement/.

Dillman, Don A., R. D. Tortora, and D. Bowker. 1998. SESRC Technical Report 98-50 *Principles for Constructing Web Surveys*. Washington.

Eutsler, Jared, and Bradley Lang. 2015. “Rating Scales in Accounting Research: The Impact of Scale Points and Labels.” *Behavioral Research in Accounting* 27(2): 35–51.

Fellegi, I. P. 1964. “Response Variance and Its Estimation.” *Journal of the American Statistical Association* 59(308): 1016–41.

Fuller, Wayne A. 1987. *Measurement Error Models*. New York: John Wiley & Sons.

Goldberger, Arthur Stanl, and Otis D. Duncan. 1973. *Structural Equation Models in the Social Sciences*. New York: Seminar Press.

Greene, Vernon L., and Edward G. Carmines. 1980. “Assessing the Reliability of Linear Composites.” *Sociological Methodology* 11: 160–75.

Groen, Jeffrey A. 2012. “Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures.” *Journal of Official Statistics* 28(2): 173–98.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. Hoboken, New Jersey: Wiley.

———. 2004. “Measurement Error Across Disciplines.” In *Measurement Error in Surveys*, eds. Paul P. Biemer et al. Hoboken, NJ, USA: John Wiley & Sons, Inc., 1–25.

———. 2009. *Survey Methodology*. Wiley.

Gustafson, Paul, and Sander Greenland. 2014. "Misclassification." In *Handbook of Epidemiology*, eds. Wolfgang Ahrens and Iris Pigeot. New York: Springer, 639–58.

Hagenaars, Jacques A. P. 1993. Sage Publications *Loglinear Models with Latent Variables*. Newbury Park, CA.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences of the United States of America* 112(8): 2395–2400.

Hansen, Morris H., William N. Hurwitz, and Max A. Bershad. 1961. "Measurement Errors in Censuses and in Surveys." *Bulletin of the International Statistical Institute* 38: 359–74.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Second. New York: Springer-Verlag.

Heise, David R. 1969. "Separating Reliability and Stability in Test-Retest Correlation." *American Sociological Review* 34(1): 91–101.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15(3): 651–74.

Japec, Lili et al. 2015. "Big Data in Survey Research: AAPOR Task Force Report." *Public Opinion Quarterly* 79(4): 839–80.

Jöreskog, Karl G. 1970. "Estimation and Testing of Simplex Models." *British Journal of Mathematical and Statistical Psychology1* 23(2): 121–45.

Jöreskog, Karl G., and Dag Sörbom. 1996. *LISREL 8: User's Reference Guide*. Uppsala, Sweden: Scientific Software International.

Jowell, Roger, Max Kaase, Rory Fitzgerald, and Gillian Eva. 2007. "The European Social Survey as a Measurement Model." In *Measuring Attitudes Cross-Nationally*, eds. Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva. London: Sage Publications, 1–29.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98(1): 191–207.

Költringer, Richard. 1995. "Measurement Quality in Austrian Personal Interview Surveys." In *The Multitrait-Multimethod Approach to Evaluate Measurement Instrument*, eds. Willem E. Saris and Ákos Münnich. Budapest: Eötvös University Press, 207–25.

Komorita, S. S., and W. K. Graham. 1965. "Number of Scale Points and the Reliability of Scales." *Educational and Psychological Measurement* 25(4): 987–95.

Krebs, Dagmar, and Juergen H.P. Hoffmeyer-Zlotnik. 2010. "Positive First or Negative First?" *Methodology* 6(3): 118–27.

Kreuter, Frauke, G. Mueller, and M. Trappmann. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74(5): 880–906.

Kreuter, Frauke, and Roger D. Peng. 2014. "Extracting Information from Big Data: Issues of Measurement, Inference and Linkage." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, eds. Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. New York: Cambridge University Press, 257–75.

Krosnick, Jon A. 1999. “Survey Research.” *Annual review of psychology* 50: 537–67.

Krosnick, Jon A., and Matthew K. Berent. 1993. “Comparisons of Party Identifications and Policy Preferences: The Impact of Survey Question Format.” *American Journal of Political Science* 37(3): 941–64.

Krosnick, Jon A., and Leandre R. Fabrigar. 1997. “Designing Rating Scales for Effective Measurement in Surveys.” In *Survey Measurement and Process Quality*, eds. Lars E. Lyberg et al. Hoboken, NJ, US, 141–64.

Krosnick, Jon A., and Stanley Presser. 2010. “Question and Questionnaire Design.” In *Handbook of Survey Research*, eds. Peter V. Marsden and James D. Write. Emerald Group Publishing Limited, 263–313.

Lazarsfeld, Paul F. 1950. “The Logical and Mathematical Foundation of Latent Structure Analysis.” In *Measurement and Prediction*, eds. Samuel A. Stouffer et al. Princeton: Princeton University Press, 362–472.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2(3): 18–22.

Lilleoja, Laur, and Willem E. Saris. 2015. “Does Correction for Measurement Error Have an Effect on the Structure and Comparability of Basic Human Values?” *Survey Research Methods* 9(3): 169–87. https://www.upf.edu/survey/_pdf/RECSM_wp045.pdf.

Loken, Eric, and Andrew Gelman. 2017. “Measurement Error and the Replication Crisis.” *Science* 355(6325): 584–85.

Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Lundmark, Sebastian, Mikael Gilljam, and Stefan Dahlberg. 2016. “Measuring Generalized Trust. An Examination of Question Wording and the Number of Scale

Points." *Public Opinion Quarterly* 80(1): 26–43.

Madans, Jennifer, Kirsten Miller, Aaron Maitland, and Gordon B. Willis. 2011. *Question Evaluation Methods*. Wiley.

Menold, Natalja, Lars Kaczmirek, Timo Lenzner, and Ales Neusar. 2014. "How Do Respondents Attend to Verbal Labels in Rating Scales?" *Field Methods* 26(1): 21–39.

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*, eds. Willem E. Saris and A. van Meurs. Amsterdam: North Holland, 134–46.

Moors, Guy, Natalia D. Kieruj, and Jeroen K. Vermunt. 2014. "The Effect of Labeling and Numbering of Response Scales on the Likelihood of Response Bias." *Sociological Methodology* 44(1): 369–99.

Oberski, Daniel L., A. Kirchner, Stephanie Eckman, and Frauke Kreuter. 2017. "Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models." *Journal of the American Statistical Association*.

Oberski, Daniel L., and Albert Satorra. 2013. "Measurement Error Models With Uncertainty About the Error Variance." *Structural Equation Modeling: A Multidisciplinary Journal* 20(3): 409–28.

Preston, Carolyn C, and Andrew M Colman. 2000. "Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences." *Acta Psychologica* 104: 1–15.

R Core Team. 2017. "R: A Language and Environment for Statistical Computing." http://www.r-project.org/.

Rammstedt, Beatrice, and Dagmar Krebs. 2007. "Does Response Scale Format Affect the Answering of Personality Scales?" *European Journal of Psychological Assessment* 23(1): 32–38.

Revilla, Melanie, and Carlos Ochoa. 2015. "Quality of Different Scales in an Online Survey in Mexico and Colombia." *Journal of Politics in Latin America* 7(3): 157–77.

Revilla, Melanie, Willem E. Saris, and Jon A. Krosnick. 2014. "Choosing the Number of Categories in Agree-Disagree Scales." *Sociological Methods & Research* 43(1): 73–97.

Revilla, Melanie, Diana Zavala-Rojas, and Willem E. Saris. 2016. "Creating a Good Question: How to Use Cumulative Experience." In *The SAGE-Handbook of Survey Methodology*, eds. Christof Wolf, Dominique Joye, Tom W. Smith, and Yang-Chih Fu. SAGE, 236–54.

Rodgers, Willard L., Frank M. Andrews, and A. Regula Herzog. 1992. "Quality of Survey Measures: A Structural Modeling Approach." *Journal of Official Statistics* 8(3): 251–75.

Saris, Willem E. et al. 2011. *The Development of the Program SQP 2.0 for the Prediction of the Quality of Survey Questions*. Barcelona, Spain. http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf.

Saris, Willem E., and Frank M. Andrews. 1991. "Evaluation of Measurement Instruments Using a Structural Modeling Approach." In *Measurement Errors in Surveys*, eds. Paul P. Biemer et al. New York: John Wiley and Sons, Inc., 575–98.

Saris, Willem E., and Irmtraud N. Gallhofer. 2007. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ, US: John Wiley and Sons, Inc.

———. 2014. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Second Ed. Hoboken, NJ, US: John Wiley and Sons, Inc.

Saris, Willem E., and Melanie Revilla. 2016. “Correction for Measurement Errors in Survey Research: Necessary and Possible.” *Social Indicators Research* 127(3): 1005–1020.

Saris, Willem E., Melanie Revilla, Jon A. Krosnick, and Eric M. Shaeffer. 2010. “Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options.” *Survey Research Methods* 4(1): 61–79.

Saris, Willem E., Albert Satorra, and Germà Coenders. 2004. “A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design.” *Sociological Methodology* 34(1): 311–47.

Scherpenzeel, Annette C., and Willem E. Saris. 1997. “The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies.” *Sociological Methods and Research* 25(3): 341–83.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. Sage Publications, Inc.

Strobl, Carolin et al. 2008. “Conditional Variable Importance for Random Forests.” *BMC Bioinformatics* 9(1): 307.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC Bioinformatics* 8(25): 21.

Sturgis, Patrick, Caroline Roberts, and Patten Smith. 2014. “Middle Alternatives Revisited: How the Neither/nor Response Acts as a Way of Saying ‘I Don’t Know’?” *Sociological Methods & Research* 43(1): 15–38.

Sudman, Seymour, and Norman M. Bradburn. 1983. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco, CA, US: Jossey Bass.

Survey Quality Predictor. 2017. Universitat Pompeu Fabra *SQP Coding Instructions*. Barcelona, Spain. http://sqp.upf.edu/media/files/sqp_coding_instructions.pdf.

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

Weijters, Bert, Elke Cabooter, and Niels Schillewaert. 2010. “The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels.” *International Journal of Research in Marketing* 27(3): 236–47.

West, B. T., and R. J. A. Little. 2013. “Nonresponse Adjustment of Survey Estimates Based on Auxiliary Variables Subject to Error.” *Journal of the Royal Statistical Society-Series C (Applied Statistics)* 62(2): 213–31.

Wiley, David E., and James A. Wiley. 1970. “The Estimation of Measurement Error in Panel Data.” *American Sociological Review* 35(1): 112–17.

Youyou, Wu, Michal Kosinski, and David Stillwell. 2015. “Computer-Based Personality Judgments Are More Accurate than Those Made by Humans.” *Proceedings of the National Academy of Sciences of the United States of America* 112(4): 1036–40.

# Appendix

**Table A1: Detailed description of the studies included**

| Study name[1] | Country[2] | Year | Mode[3] | Design[4] | Questions | Sample |
|---|---|---|---|---|---|---|
| IFES 1992: Political parties | Austria | 1992 | FTF | 1x4x3 | 12 | 877 |
| IFES 1992: Economic expectations | Austria | 1992 | FTF | 1x4x3 | 12 | 1149 |
| IFES 1992: Materialism/ Postmaterialism | Austria | 1992 | FTF | 1x4x3 | 12 | 1404 |
| IFES 1992: Life management | Austria | 1992 | FTF | 1x4x3 | 12 | 956 |
| IFES 1992: European Community | Austria | 1992 | FTF | 1x4x4 | 16 | 1486 |
| IFES 1992: Subjective life satisfaction 1 | Austria | 1992 | FTF | 1x4x3 | 12 | 1666 |
| IFES 1992: Subjective life satisfaction 2 | Austria | 1992 | FTF | 1x4x3 | 12 | 1627 |
| IFES 1992: Trust in institutions 1 | Austria | 1992 | FTF | 1x3x3 | 9 | 1553 |
| IFES 1992: Trust in institutions 2 | Austria | 1992 | FTF | 1x3x3 | 9 | 1522 |
| IFES 1992: Physical well-being | Austria | 1992 | FTF | 1x4x3 | 12 | 1601 |
| KUL 1989: Life satisfaction | Belgium | 1989 | FTF | 1x5x3 | 15 | 1605 |
| KUL 1997: Immigration Threat | Belgium | 1997 | FTF/Mail | 1x3x3 | 9 | 542 |
| KUL 1997: Immigration Outgroup | Belgium | 1997 | FTF/Mail | 1x3x3 | 9 | 542 |

[1]The questions and the results of the analysis of these experiments can be found in the public database of the second version of the Survey Quality Predictor (SQP) by searching for the study name indicated in the table.
[2]List of countries' ISO code by ESS study:
ESS Round 1: AT, BE, CH, CZ, DE, DK, ES, FI, FR, GB, GR, IE, NL, NO, PL, PT, SE, SI
ESS Round 2: AT, BE, CH, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, NL, NO, PL, PT, SI, SK, UA
ESS Round 3: AT, BE, CH, DE, DK, EE, ES, FI, FR, GB, IE, NL, NO, PL, PT, SI, SK, UA
[3]FTF stands for face to face interviewing, Mail stands for self-completion questionnaires send by ordinal post mailing, Telephone stands for computer assisted telephone interviewing (CATI) and Telepanel is a type of computer assisted self-administered interviews (CASI) very popular in the 90's where the respondents were provided with a computer and a modem when necessary. In the ESS studies the mode of data collection is usually FTF although in some countries the supplementary questionnaire is self-completed.
[4]The design of the experiments consists of *n* experiments, *t* traits and *m* methods: *n* x *t* x *m*

| | | | | | | |
|---|---|---|---|---|---|---|
| KUL 1997: Immigration Ingroup | Belgium | 1997 | FTF/Mail | 1x4x3 | 12 | 542 |
| STP 1992: Seriousness of crimes | Netherlands | 1992 | Mail/telepanel | 2x3x2 | 12 | 674 |
| STP 1991: Political efficacy (Europe) | Netherlands | 1991 | Telepanel | 2x4x2 | 16 | 915 |
| NIMMO 1992: European politics | Netherlands | 1992 | Mail/telepanel | 2x3x2 | 12 | 601 |
| NIMMO 1992: Life satisfaction 1 | Netherlands | 1992 | Telephone | 2x4x2 | 16 | 915 |
| NIMMO 1991: Life satisfaction 2 | Netherlands | 1991 | Mail | 2x4x2 | 16 | 456 |
| NIMMO 1992: Life satisfaction 3 | Netherlands | 1992 | Mail | 2x4x2 | 16 | 456 |
| NIMMO/STP 1992: Life satisfaction 4 | Netherlands | 1992 | Mail/telepanel | 2x4x2 | 16 | 363 |
| NIPO 1989: Life satisfaction 5 | Netherlands | 1989 | Telepanel | 1x4x3 | 12 | 486 |
| STP 1991: Life satisfaction 6 | Netherlands | 1991 | Telepanel | 2x4x2 | 16 | 1143 |
| STP 1991: Life satisfaction 7 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 1599 |
| STP 1992: Values in life | Netherlands | 1992 | Mail/telepanel | 2x3x2 | 12 | 677 |
| STP 1991: Values of comfort/self-respect/status 1 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 368 |
| STP 1991: Values of family/ ambition/independence 1 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 368 |
| STP 1991: Values of comfort/self-respect/status 2 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 408 |
| STP 1991: Values of family/ ambition/independence 2 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 408 |
| STP 1991: Values of comfort/self-respect/status 3 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 390 |
| STP 1991: Values of family/ ambition/independence 3 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 390 |
| STP 1991: Values of comfort/self-respect/status 4 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 422 |
| STP 1991: Values of family/ ambition/independence 4 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 422 |
| STP 1991: Seriousness of crimes 1 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 627 |
| STP 1991: Seriousness of crimes 2 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 627 |
| STP 1991: Seriousness of crimes 3 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 627 |
| STP 1991: Seriousness of crimes 4 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 631 |

| | | | | | | |
|---|---|---|---|---|---|---|
| STP 1991: Seriousness of crimes 5 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 631 |
| STP 1991: Seriousness of crimes 6 | Netherlands | 1991 | Telepanel | 2x3x2 | 12 | 631 |
| STP 1991: EU membership | Netherlands | 1991 | Telepanel | 2x4x2 | 16 | 927 |
| STP 1991: Crimes 1 | Netherlands | 1991 | Telepanel | 1x3x3 | 9 | 344 |
| STP 1991: Crimes 2 | Netherlands | 1991 | Telepanel | 1x3x3 | 9 | 344 |
| STP 1991: Crimes 3 | Netherlands | 1991 | Telepanel | 1x3x3 | 9 | 344 |
| NIPO 1988: Trade-unions 1 | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 438 |
| NIPO 1988: Trade-unions 2 | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 416 |
| NIPO 1988: Trade-unions 3 | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 403 |
| NIPO 1988: Trade-unions 4 | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 398 |
| STP 1992: Life satisfaction 1 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 216 |
| STP 1992: Life satisfaction 2 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 252 |
| STP 1992: Life satisfaction 3 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 248 |
| STP 1992: Life satisfaction 4 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 244 |
| STP 1992: Work conditions 1 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 361 |
| STP 1992: Work conditions 2 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 420 |
| STP 1992: Work conditions 3 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 388 |
| STP 1992: Work conditions 4 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 389 |
| STP 1992: Living conditions 1 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 219 |
| STP 1992: Living conditions 2 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 253 |
| STP 1992: Living conditions 3 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 249 |
| STP 1992: Living conditions 4 | Netherlands | 1992 | Telepanel | 2x3x2 | 12 | 245 |
| STP 1988: Evaluating TV programs | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 575 |
| STP 1988: Use of the TV | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 578 |
| STP 1988: Reading newspaper | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 578 |
| STP 1988: Evaluation of politics | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 544 |
| STP 1988: Evaluation of politicians | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 544 |

| | | | | | | |
|---|---|---|---|---|---|---|
| STP 1988: Political participation | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 554 |
| STP 1988: Estimation of income | Netherlands | 1988 | Telepanel | 1x3x3 | 9 | 554 |
| STP 1996: Social trust | Netherlands | 1996 | Telepanel | 2x4x2 | 16 | 2444 |
| STP 1996: Social problems | Netherlands | 1996 | Telepanel | 2x4x2 | 16 | 2444 |
| STP 1996: Ethnic minorities 1 | Netherlands | 1996 | Telepanel | 2x3x2 | 12 | 2444 |
| STP 1996: Ethnic minorities 2 | Netherlands | 1996 | Telepanel | 2x4x2 | 16 | 2444 |
| STP 1996: Ethnic minorities 3 | Netherlands | 1996 | Telepanel | 2x4x2 | 16 | 2444 |
| SRC January 1979: Finances, business, health, news 1 | United States | 1979 | FTF | 1x4x3 | 12 | 229 |
| SRC January 1979: Finances, business, health, news 2 | United States | 1979 | FTF | 1x4x3 | 12 | 229 |
| SRC January 1979: Finances, business, health, news 3 | United States | 1979 | FTF | 1x4x3 | 12 | 227 |
| SRC January 1979: Finances, business, health, news 4 | United States | 1979 | FTF | 1x4x3 | 12 | 227 |
| SRC July 1981: Finances, business, health, last year | United States | 1981 | FTF | 1x3x3 | 9 | 703 |
| SRC July 1981: Finances, business, health, next year | United States | 1981 | FTF | 1x3x3 | 9 | 703 |
| SRC 1981: Satisfaction with life | United States | 1981 | FTF | 1x4x3 | 12 | 866 |
| MGP 1984-1987: Health and income 1 | United States | 1986 | FTF | 2x2x3 | 12 | 901 |
| MGP 1984-1987: Saving/ transport/safety | United States | 1986 | FTF | 2x3x2 | 12 | 893 |
| MGP 1984-1987: Feelings 1 | United States | 1986 | FTF | 2x3x3 | 18 | 1197 |
| MGP 1984-1987: Feelings 2 | United States | 1986 | FTF | 2x3x3 | 18 | 1197 |
| MGP 1984-1987: Health and income 2 | United States | 1986 | FTF | 2x4x2 | 16 | 1169 |
| MGP 1984-1987: Life in general | United States | 1986 | FTF | 2x5x2 | 20 | 867 |
| ESS Round 1: Social trust | 18 | 2002 | FTF | 1x3x3 | 156 | |
| ESS Round 1: Internal political efficacy | 18 | 2002 | FTF | 1x3x3 | 159 | 36,065 |
| ESS Round 1: Trust in institutions | 18 | 2002 | FTF | 1x3x3 | 156 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| ESS Round 1: Political satisfaction | 18 | 2002 | FTF | 1x3x3 | 156 | |
| ESS Round 1: Evaluation of policies | 18 | 2002 | FTF | 1x3x3 | 156 | |
| ESS Round 2: Trust in institutions | 18 | 2004 | FTF | 1x3x3 | 162 | |
| ESS Round 2: Political satisfaction | 20 | 2004 | FTF | 1x3x3 | 180 | |
| ESS Round 2: Doctors social distance | 18 | 2004 | FTF | 1x3x3 | 162 | 38,643 |
| ESS Round 2: Parents role | 18 | 2004 | FTF | 1x3x3 | 162 | |
| ESS Round 2: Evaluation of job | 1 | 2004 | FTF | 1x3x3 | 171 | |
| ESS Round 3: Attitudes towards immigration | 18 | 2006 | FTF | 1x3x4 | 216 | |
| ESS Round 3: Evaluation of immigration | 18 | 2006 | FTF | 1x3x4 | 216 | 34,577 |
| ESS Round 3: Personal accomplishment | 18 | 2006 | FTF | 1x3x4 | 216 | |
| ESS Round 3: Social value of life | 16 | 2006 | FTF | 1x3x4 | 192 | |

**Table A2: List of predictor variables included in the predictive model**

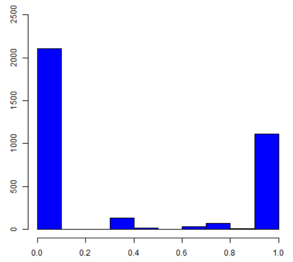
| Characteristic name | Variable name | Type of variable | Values | Labels | N.Obs. |
|---|---|---|---|---|---|
| Absolute or comparative judgment | absolute | Dummy | 0 | Absolute | 3,125 |
| | | | 1 | Comparative | 358 |
| Balance of the request | balance | Dummy | 0 | Balanced | 2,623 |
| | | | 1 | Unbalanced | 860 |
| Centrality | centrality | Categorical | 0 | Not at all central | 954 |
| | | | 1 | A bit central | 940 |
| | | | 2 | Rather central | 986 |
| | | | 3 | Central | 366 |
| | | | 4 | Very central | 237 |
| Computer assisted | computer.assisted | Dummy | 0 | No | 1,552 |
| | | | 1 | Yes | 1,931 |
| Concept | concept | Categorical | 1 | Evaluative belief | 732 |
| | | | 2 | Feeling | 1,093 |
| | | | 3 | Importance of something | 96 |

| | | | | | |
|---|---|---|---|---|---|
| | | | 4 | Expectation of future events | 39 |
| | | | 5 | Facts, background and behaviour | 111 |
| | | | 6 | Complex concepts | 44 |
| | | | 71 | Judgement | 123 |
| | | | 72 | Relationship | 8 |
| | | | 73 | Evaluation | 699 |
| | | | 74 | Preference | 2 |
| | | | 75 | Norm | 159 |
| | | | 76 | Policy | 267 |
| | | | 77 | Right | 57 |
| | | | 78 | Action tendency | 53 |
| Country | country | Categorical | AT | Austria | 256 |
| | | | BE | Belgium | 156 |
| | | | CH | Switzerland | 138 |
| | | | CZ | Czech Republic | 90 |
| | | | DE | Germany | 138 |
| | | | DK | Denmark | 138 |
| | | | EE | Estonia | 93 |
| | | | ES | Spain | 138 |
| | | | FI | Finland | 138 |
| | | | FR | France | 138 |
| | | | GB | Great Britain | 138 |
| | | | GR | Greece | 90 |
| | | | IE | Ireland | 138 |
| | | | NL | Netherlands | 824 |
| | | | NO | Norway | 102 |
| | | | PL | Poland | 138 |
| | | | PT | Portugal | 138 |
| | | | SE | Sweden | 45 |
| | | | SI | Slovenia | 138 |
| | | | SK | Slovakia | 81 |
| | | | UA | Ukraine | 54 |
| | | | US | United States | 174 |

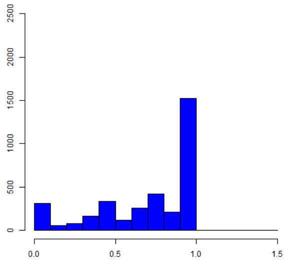
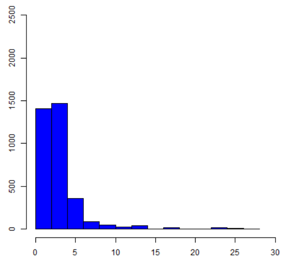
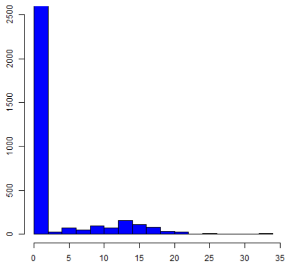
| | | | | | |
|---|---|---|---|---|---|
| Domain | domain | Categorical | 2 | International politics | 64 |
| | | | 3 | Health | 244 |
| | | | 4 | Living conditions | 433 |
| | | | 6 | Other beliefs | 292 |
| | | | 7 | Work | 516 |
| | | | 10 | Family | 109 |
| | | | 11 | Personal relations | 329 |
| | | | 13 | Consumer behaviour | 26 |
| | | | 14 | Leisure activities | 131 |
| | | | 101 | National government | 184 |
| | | | 103 | National institutions | 440 |
| | | | 105 | Political parties | 30 |
| | | | 107 | Trade unions | 12 |
| | | | 112 | Economic/financial matters | 287 |
| | | | 120 | Other | 386 |
| Don't know option | dont_know | Categorical | 1 | Present | 187 |
| | | | 2 | Only registered | 1,759 |
| | | | 3 | Not present | 1,537 |
| Presence of encouragement to answer | encourage | Dummy | 0 | Not present | 2,859 |
| | | | 1 | Present | 624 |
| Formulation of the request for an answer: basic choice | form_basic | Categorical | 0 | No request present | 1,054 |
| | | | 1 | Indirect request | 965 |
| | | | 2 | Direct request | 1,464 |
| MTMM estimate survey agency | from | Dummy | 0 | ESS | 2,460 |
| | | | 1 | Other | 1,023 |
| Reference period: Future | future | Dummy | 0 | Not future | 2,905 |
| | | | 1 | Future | 578 |
| Use of gradation | gradation | Dummy | 0 | No gradation | 1,392 |
| | | | 1 | Gradation used | 2,091 |
| Interviewer instruction | instr_interv | Dummy | 0 | Absent | 2,691 |
| | | | 1 | Present | 792 |

| | | | | | |
|---|---|---|---|---|---|
| Respondent instruction | instr_respon | Dummy | 0 | Absent | 1,897 |
| | | | 1 | Present | 1,586 |
| Interviewer | interviewer | Dummy | 0 | No | 1,087 |
| | | | 1 | Yes | 2,396 |
| Request present in the introduction | intr_request | Dummy | 0 | Not present | 3,371 |
| | | | 1 | Present | 112 |
| Introduction available? | intropresent | Dummy | 0 | Not present | 2,753 |
| | | | 1 | Present | 730 |
| Knowledge provided | knowledge | Categorical | 0 | Not provided | 2,927 |
| | | | 1 | Definitions only | 372 |
| | | | 2 | Other explanations | 163 |
| | | | 3 | Both definitions and other explan. | 21 |
| Labels of categories | labels | Categorical | 1 | No labels | 192 |
| | | | 2 | Partially labelled | 1,612 |
| | | | 3 | Fully labelled | 1,679 |
| Labels with short text or complete sentences | labels_gramm | Dummy | 0 | Short text | 3,201 |
| | | | 1 | Complete sentences | 282 |
| Order of the labels | labels_order | Categorical | 1 | First label negative | 1,993 |
| | | | 2 | First label positive | 1,490 |
| Extra information or definition | motivation | Dummy | 0 | Absent | 2,972 |
| | | | 1 | Present | 511 |
| Information about the opinion of other people | opinionother | Dummy | 0 | Absent | 2,287 |
| | | | 1 | Present | 1,196 |
| Reference period: Past | past | Dummy | 0 | Not past | 3,267 |
| | | | 1 | Past | 216 |
| Request for an answer type | questiontype | Categorical | 0 | None | 1,225 |
| | | | 1 | Interrogative | 1,605 |
| | | | 2 | Imperative | 486 |
| | | | 3 | Declarative | 167 |
| Consistency between theoretical range and used range of scale | range.correspondence | Categorical | 1 | Not applicable | 819 |
| | | | 2 | Low consistency | 1,268 |
| | | | 3 | High consistency | 1,396 |

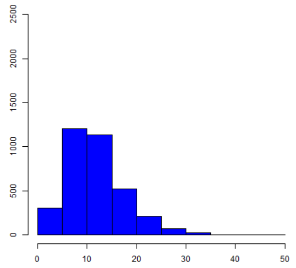
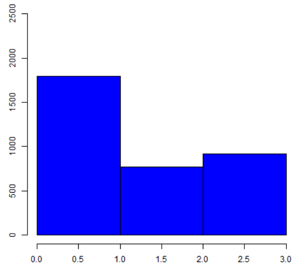
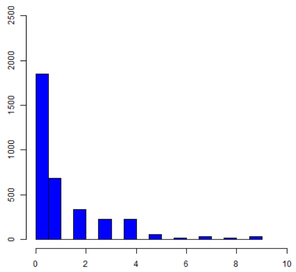
| | | | | | |
|---|---|---|---|---|---|
| Repetition | repetition | Dummy | 0 | No repetition | 2,003 |
| | | | 1 | Repetition | 1,480 |
| Neutral category | scal_neutral | Categorical | 1 | Present | 2,340 |
| | | | 3 | Not present | 1,143 |
| Response scale: basic choice | scale_basic | Categorical | 0 | More than 3 category scales | 2,922 |
| | | | 1 | Two-category scales | 88 |
| | | | 2 | Numerical open-ended answers | 77 |
| | | | 3 | Magnitude estimation | 169 |
| | | | 4 | Line production | 201 |
| | | | 5 | More steps procedures | 26 |
| Correspondence between labels and numbers of the scale | scale_corres | Categorical | 0 | Not applicable | 312 |
| | | | 1 | High correspondence | 1,326 |
| | | | 2 | Medium correspondence | 909 |
| | | | 3 | Low correspondence | 936 |
| Theoretical range of the scale bipolar/ unipolar | scale_trange | Dummy | 0 | Theoretically unipolar | 819 |
| | | | 1 | Theoretically bipolar | 2,664 |
| Range of the used scale bipolar/ unipolar | scale_urange | Dummy | 0 | Unipolar | 1,362 |
| | | | 1 | Bipolar | 2,121 |
| Scale with only numbers or numbers in boxes | showc_boxes | Dummy | 0 | Only numbers | 3,175 |
| | | | 1 | Numbers in boxes | 308 |
| Horizontal or vertical scale | showc_horiz | Dummy | 0 | Vertical | 2,467 |
| | | | 1 | Horizontal | 1,016 |
| Numbers or letters before the answer categories | showc_letter | Dummy | 0 | Neither | 2,131 |
| | | | 1 | Numbers | 1,352 |
| Overlap of scale labels and categories | showc_over | Dummy | 0 | Present | 1,913 |
| | | | 1 | Absent | 1,570 |
| Request on the visual aid | showc_quest | Dummy | 0 | No | 3,197 |
| | | | 1 | Yes | 286 |
| Start of the response sentence on the visual aid | showc_start | Dummy | 0 | No | 3,272 |
| | | | 1 | Yes | 211 |

| | | | | | |
|---|---|---|---|---|---|
| Social Desirability | socdesir | Categorical | 0 | Not present | 604 |
| | | | 1 | A bit | 2,555 |
| | | | 2 | A lot | 324 |
| Use of stimulus or statement in the request | stimulus | Dummy | 0 | Not present | 1,380 |
| | | | 1 | Present | 2,103 |
| Emphasis on subjective opinion in request | subjectiveop | Dummy | 0 | Not present | 2,934 |
| | | | 1 | Present | 549 |
| Symmetry of response scale | symmetry | Dummy | 0 | Asymmetric | 1,277 |
| | | | 1 | Symmetric | 2,206 |
| WH word used in the request | used_WH_word | Dummy | 0 | Without WH word | 2,524 |
| | | | 1 | WH word used | 959 |
| Showcard or other visual aids used | usedshowcard | Dummy | 0 | Not used | 1,583 |
| | | | 1 | Used | 1,900 |
| Visual or oral presentation | visual | Dummy | 0 | Oral | 1,787 |
| | | | 1 | Visual | 1,696 |
| Average abstract nouns in introduction | avgabs_intro[5] | Numeric | |  | |

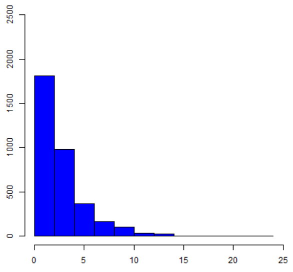
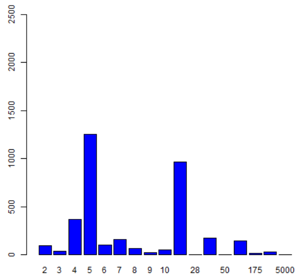
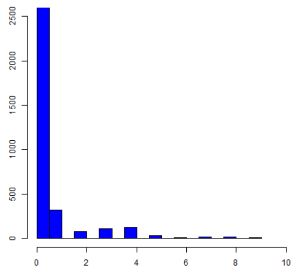
[5]This characteristic is set to a median value because the number of abstract nouns in the introduction it is not coded. Formula used: nd$avgabs_intro <- ifelse(nd$intropresent == 1, 0.67, 0).

| | | | |
|---|---|---|---|
| Total average number of abstract nouns in question | avgabs_total[6] | Numeric |  |
| Total average number of syllables in question | avgsy_total[7] | Numeric |  |
| Average number of words in introduction | avgwrd_intro[8] | Numeric |  |

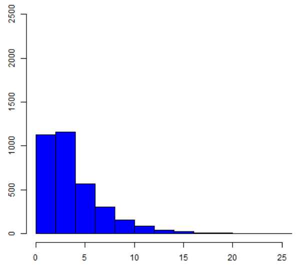
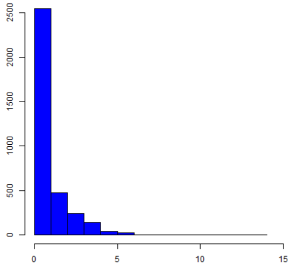
[6]Indicates the average number of abstract nouns in the question, in total, in relation to the question number of nouns. Formula used: nabst_total / nnouns_total.
[7]Indicates the average number of syllables in the question, in total, in relation to the question number of words. Formula used: nsyll_total / nwords_total.
[8]Indicates the average number of words in the introduction in relation to the number of sentences. Formula used: nwords_intro / nsents_intro.

| | | | |
|---|---|---|---|
| Total average number of words in request for an answer | avgwrd_total[9] | Numeric |  |
| Number of fixed reference points | fixrefpoints | Numeric |  |
| Number of abstract nouns in introduction | nabst_intro | Numeric |  |

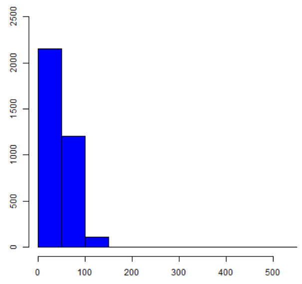
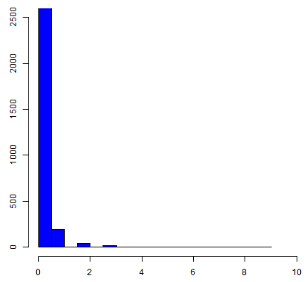
[9]Indicates the total average number of words in the request for an answer in relation to the number of sentences. Formula used: nwords_total / nsents_quest

| | | | |
|---|---|---|---|
| Total number of abstract nouns in question | nabst_total[10] | Numeric |  |
| Number of categories | ncategories | Numeric |  |
| Number of nouns in introduction | nnouns_intro | Numeric |  |

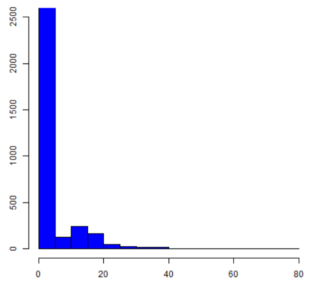
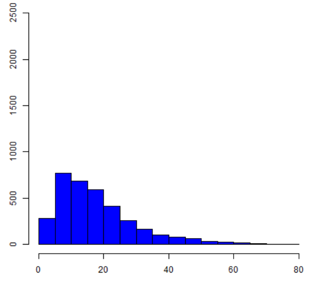
[10] It refers to the sum of the number of abstract nouns in the request and the number of abstract nouns in the answer scale. Formula used: nabst_quest + nabst_ans.

| | | | | |
|---|---|---|---|---|
| Total number of nouns in question | nnouns_total[11] | Numeric |  | |
| Number of subordinate clauses in request | nsub_quest | Numeric |  | |

[11]Indicates the sum of the number of nouns in the request for an answer and the nouns in the answer scale. Formula used: nnouns_quest + nnouns_ans.

| | | | |
|---|---|---|---|
| Total number of syllables in question | nsyll_total[12] | Numeric |   |
| Number of subordinated clauses in introduction | numsub_intro | Numeric |   |

[12]Indicates the sum of the number of syllables in the request for an answer and the syllables in the answer options. Formula used: nsyll_quest + nsyll_ans.

| | | | |
|---|---|---|---|
| Number of words in introduction | nwords_intro | Numeric |  |
| Total number of words in request | nwords_total[13] | Numeric |  |

[13]It refers to the number of words of the request for an answer text, because the number of words in the answer options is not coded.

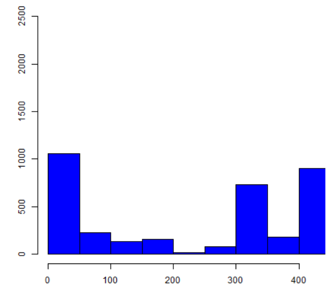
| | | | | |
|---|---|---|---|---|
| Position | position | Numeric |  | |

**Table A3: Complete list of predictors by highest marginal permutation importance rom the reduced random forest prediction model**

| Rank | Logit Reliability coefficient | | | Logit Method effect | | |
|---|---|---|---|---|---|---|
| | **Predictor** | **Question characteristic** | **Propensity MSE** | **Predictor** | **Question characteristic** | **Propensity MSE** |
| 1 | domain | Topic of the question | 0.208 | ncategories | No. categories | 0.130 |
| 2 | ncategories | No. categories | 0.165 | position | Position in q'naire | 0.116 |
| 3 | country | Country of administration | 0.095 | domain | Topic of the question | 0.113 |
| 4 | concept | Concept being measured | 0.086 | country | Country of administration | 0.070 |
| 5 | position | Position in q'naire | 0.058 | concept | Concept being measured | 0.067 |
| 6 | nsyll_total | No. syllables total | 0.035 | labels | Partial or full labels | 0.045 |
| 7 | avgwrd_total | Words/sentences total | 0.028 | avgwrd_total | Words/sentences total | 0.028 |
| 8 | avgsy_total | Syllables/words total | 0.021 | scale_basic | Type answer scale | 0.024 |
| 9 | nwords_total | No. words total | 0.021 | repetition | MTMM repetition | 0.022 |
| 10 | labels | Partial or full labels | 0.017 | scale_corres | Numbering agree with labels | 0.021 |
| 11 | scale_corres | Numbers agree with labels | 0.015 | nwords_total | No. words total | 0.019 |
| 12 | nnouns_total | No. nouns total | 0.014 | avgsy_total | Syllables/words total | 0.019 |
| 13 | nabst_total | No. abstract nouns total | 0.013 | nnouns_total | No. nouns total | 0.019 |
| 14 | avgabs_total | Abstract nouns/nouns total | 0.012 | nsyll_total | No. syllables total | 0.018 |
| 15 | questiontype | Grammatical form | 0.012 | range.corres | Polarity correspondence | 0.016 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16 | centrality | Centrality of topic | 0.011 | avgabs_total | Abstract nouns/nouns total | 0.014 |
| 17 | repetition | MTMM repetition | 0.010 | nsub_quest | No. subordinate clauses request | 0.013 |
| 18 | nsub_quest | No. subordinate clauses request | 0.010 | labels_order | Negative/positive label order | 0.011 |
| 19 | fixrefpoints | No. fixed reference points | 0.010 | centrality | Centrality of topic | 0.011 |
| 20 | nabst_intro | No. abstract nouns intro | 0.009 | nabst_total | No. abstract nouns total | 0.011 |
| 21 | range.corres | Polarity correspondence | 0.009 | questiontype | Grammatical form | 0.011 |
| 22 | nnouns_intro | No. nouns intro | 0.008 | dont_know | Use of don’t know option | 0.011 |
| 23 | labels_order | Negative/positive label order | 0.007 | stimulus | Item in battery | 0.010 |
| 24 | avgwrd_intro | Words/sentences intro | 0.007 | fixrefpoints | No. fixed reference points | 0.010 |
| 25 | form_basic | Formulation request | 0.006 | form_basic | Formulation request | 0.009 |
| 26 | dont_know | Use of don’t know option | 0.006 | knowledge | Type of knowledge | 0.009 |
| 27 | scale_basic | Type of answer scale | 0.006 | from | MTMM estimate survey agency | 0.008 |
| 28 | avgabs_intro | Abstract nouns/nouns intro | 0.005 | nabst_intro | No. abstract nouns intro | 0.008 |
| 29 | visual | Oral or visual presentation | 0.005 | visual | Oral or visual presentation | 0.008 |
| 30 | from | MTMM estimate survey agency | 0.005 | balance | Balance of the request | 0.007 |
| 31 | nwords_intro | No. words intro | 0.005 | scale_urange | Polarity used in scale | 0.007 |
| 32 | future | Reference period: Future | 0.005 | interviewer | Interviewer presence | 0.007 |
| 33 | socdesir | Social desirability | 0.005 | nwords_intro | No. words intro | 0.007 |
| 34 | stimulus | Item in battery | 0.004 | nnouns_intro | No. nouns intro | 0.006 |
| 35 | scale_trange | Theoretical polarity concept | 0.003 | socdesir | Social desirability | 0.006 |
| 36 | instr_respon | Respondent instruction | 0.003 | showc_letter | Use of numbers or letters in showcard | 0.005 |
| 37 | balance | Balance of the request | 0.003 | motivation | Extra information | 0.005 |
| 38 | gradation | Use of gradation | 0.003 | symmetry | Scales’ symmetry | 0.005 |
| 39 | opinionother | Info. opinion other people | 0.003 | avgwrd_intro | Words/sentences intro | 0.005 |
| 40 | computer.assisted | Computer administration | 0.003 | avgabs_intro | Abstract nouns/nouns intro | 0.005 |
| 41 | scal_neutral | Use of neutral category | 0.003 | instr_respon | Respondent instruction | 0.004 |
| 42 | motivation | Extra information | 0.003 | showc_horiz | Horizontal layout in showcard | 0.004 |
| 43 | encourage | Presence of encouragement | 0.003 | showc_over | Overlap in showcard | 0.004 |
| 44 | knowledge | Type of knowledge | 0.003 | gradation | Use of gradation | 0.004 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 45 | interviewer | Interviewer presence | 0.002 | scal_neutral | Use of neutral category | 0.004 |
| 46 | symmetry | Scales' symmetry | 0.002 | used_WH_word | Use of WH word | 0.004 |
| 47 | subjectiveop | Emphasis subjective opinion | 0.002 | computer.assisted | Computer administration | 0.003 |
| 48 | used_WH_word | Use of WH word | 0.002 | instr_interv | Interviewer instruction | 0.003 |
| 49 | scale_urange | Polarity used in scale | 0.002 | encourage | Presence of encouragement | 0.003 |
| 50 | showc_letter | Use of numbers or letters in showcard | 0.002 | opinionother | Info. opinion other people | 0.003 |
| 51 | showc_over | Overlap in showcard | 0.002 | future | Reference period: future | 0.003 |
| 52 | showc_boxes | Use of boxes in showcard | 0.002 | showc_quest | Request in showcard | 0.003 |
| 53 | instr_interv | Interview instruction | 0.002 | scale_trange | Theoretical polarity concept | 0.003 |
| 54 | showc_horiz | Horizontal layout in showcard | 0.002 | absolute | Absolute or comparative judgement | 0.003 |
| 55 | absolute | Absolute or comparative judgement | 0.002 | used_showcard | Use of showcard | 0.002 |
| 56 | showc_quest | Request in showcard | 0.002 | subjectiveop | Emphasis subjective opinion | 0.002 |
| 57 | numsub_intro | No. subordinate clauses intro | 0.002 | numsub_intro | No. subordinate clauses intro | 0.002 |
| 58 | usedshowcard | Use of showcard | 0.002 | showc_boxes | Use of boxes in showcard | 0.002 |
| 59 | intropresent | Use of introduction | 0.002 | labels_gramm | Labels' length | 0.002 |
| 60 | labels_gramm | Labels' length | 0.001 | intr_request | Request in introduction | 0.002 |
| 61 | past | Reference period: Past | 0.001 | showc_start | Start of response sentence in showcard | 0.002 |
| 62 | showc_start | Start of response sentence in showcard | 0.001 | past | Reference period: Past | 0.001 |
| 63 | intr_request | Request in introduction | 0.001 | intropresent | Use of introduction | 0.001 |

**Table A4: Complete list of predictors by highest conditional permutation importance from the reduced random forest prediction model**

| Rank | Logit Reliability coefficient | | | Logit Method effect | | |
|---|---|---|---|---|---|---|
| | **Predictor** | **Question characteristic** | **Accuracy decrease** | **Predictor** | **Question characteristic** | **Accuracy decrease** |
| 1 | ncategories | No. categories | 0.128 | ncategories | No. categories | 0.170 |
| 2 | scale_corres | Numbers agree with labels | 0.052 | labels | Partial or full labels | 0.121 |
| 3 | labels | Partial or full labels | 0.051 | position | Position in q'naire | 0.103 |
| 4 | questiontype | Grammatical form | 0.046 | scale_corres | Numbers agree with labels | 0.066 |
| 5 | labels_order | Negative/positive label order | 0.032 | range.corres | Polarity correspondence | 0.058 |
| 6 | range.corres | Polarity correspondence | 0.032 | visual | Oral or visual presentation | 0.057 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | avgwrd_total | Words/sentences total | 0.031 | scale_basic | Type of answer scale | 0.052 |
| 8 | stimulus | Item in battery | 0.027 | labels_order | Negative/positive label order | 0.038 |
| 9 | scale_basic | Type of answer scale | 0.021 | stimulus | Item in battery | 0.037 |
| 10 | fixrefpoints | No. fixed reference points | 0.021 | showc_horiz | Horizontal layout in showcard | 0.034 |
| 11 | position | Position in q'naire | 0.020 | scale_urange | Polarity used in scale | 0.034 |
| 12 | form_basic | Formulation request | 0.018 | symmetry | Scales’ symmetry | 0.028 |
| 13 | interviewer | Interviewer presence | 0.018 | avgwrd_total | Words/sentences total | 0.028 |
| 14 | gradation | Use of gradation | 0.017 | avgabs_total | Abstract nouns/nouns total | 0.024 |
| 15 | nsyll_total | No. syllables total | 0.017 | form_basic | Formulation request | 0.022 |
| 16 | showc_over | Overlap in showcard | 0.016 | balance | Balance of request | 0.021 |
| 17 | scal_neutral | Use of neutral category | 0.015 | dont_know | Use of don’t know option | 0.020 |
| 18 | opinionother | Info. opinion other people | 0.015 | questiontype | Grammatical form | 0.020 |
| 19 | balance | Balance of request | 0.014 | showc_over | Overlap in showcard | 0.015 |
| 20 | nwords_total | No. words total | 0.014 | scal_neutral | Use of neutral category | 0.014 |
| 21 | avgabs_total | Abstract nouns/nouns total | 0.013 | avgsy_total | Syllables/words total | 0.014 |
| 22 | avgsy_total | Syllables/words total | 0.013 | interviewer | Interviewer presence | 0.014 |
| 23 | scale_urange | Polarity used in scale | 0.012 | usedshowcard | Use of showcard | 0.014 |
| 24 | visual | Oral or visual presentation | 0.011 | nwords_total | No. words total | 0.014 |
| 25 | scale_trange | Theoretical polarity concept | 0.011 | scale_trange | Theoretical polarity concept | 0.013 |
| 26 | showc_horiz | Horizontal layout in showcard | 0.010 | showc_letter | Use of numbers or letters in showcard | 0.013 |
| 27 | used_WH_word | Use of WH word | 0.009 | nnouns_total | No. nouns total | 0.013 |
| 28 | computer.assisted | Computer administration | 0.009 | knowledge | Type of knowledge | 0.013 |
| 29 | symmetry | Scales’ symmetry | 0.009 | nsyll_total | No. syllables total | 0.013 |
| 30 | knowledge | Type of knowledge | 0.008 | intropresent | Use of introduction | 0.012 |
| 31 | dont_know | Use of don’t know option | 0.008 | gradation | Use of gradation | 0.012 |
| 32 | nnouns_total | No. nouns total | 0.008 | instr_respon | Respondent instruction | 0.011 |
| 33 | usedshowcard | Use of showcard | 0.007 | motivation | Extra information | 0.011 |
| 34 | encourage | Presence of encouragement | 0.006 | fixrefpoints | No. fixed reference points | 0.010 |
| 35 | intropresent | Use of introduction | 0.006 | nsub_quest | No. subordinate clauses request | 0.010 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 36 | nsub_quest | No. subordinate clauses request | 0.006 | opinionother | Info. opinion other people | 0.009 |
| 37 | intr_request | Request in introduction | 0.005 | absolute | Absolute or comparative judgement | 0.009 |
| 38 | motivation | Extra information | 0.005 | used_WH_word | Use of WH word | 0.008 |
| 39 | nnouns_intro | No. nouns intro | 0.005 | avgwrd_intro | Words/sentences intro | 0.007 |
| 40 | avgwrd_intro | Words/sentences intro | 0.005 | instr_interv | Interviewer instruction | 0.006 |
| 41 | avgabs_intro | Abstract nouns/nouns intro | 0.005 | subjectiveop | Emphasis subjective opinion | 0.006 |
| 42 | subjectiveop | Emphasis subjective opinion | 0.005 | avgabs_intro | Abstract nouns/nouns intro | 0.006 |
| 43 | absolute | Absolute or comparative judgement | 0.005 | computer.assisted | Computer administration | 0.006 |
| 44 | nabst_intro | No. abstract nouns intro | 0.005 | nnouns_intro | No. nouns intro | 0.005 |
| 45 | nabst_total | No. abstract nouns total | 0.005 | encourage | Presence of encouragement | 0.005 |
| 46 | instr_respon | Respondent instruction | 0.004 | nwords_intro | No. words intro | 0.005 |
| 47 | nwords_intro | No. words intro | 0.004 | nabst_total | No. abstract nouns total | 0.004 |
| 48 | labels_gramm | Labels' length | 0.004 | nabst_intro | No. abstract nouns intro | 0.004 |
| 49 | showc_letter | Use of numbers or letters in showcard | 0.003 | intr_request | Request in introduction | 0.003 |
| 50 | instr_interv | Interviewer instruction | 0.002 | showc_quest | Request in showcard | 0.003 |
| 51 | showc_start | Start of response sentence in showcard | 0.002 | labels_gramm | Labels' length | 0.002 |
| 52 | showc_quest | Request in showcard | 0.002 | numsub_intro | No. subordinate clauses intro | 0.001 |
| 53 | numsub_intro | No. subordinate clauses intro | 0.001 | showc_start | Start of response sentence in showcard | 0.001 |
| 54 | showc_boxes | Use of boxes in showcard | 0.001 | showc_boxes | Use of boxes in showcard | 0.001 |