mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology
Volume 14, 2020 | 2

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.



Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246526
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.
All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Layout: Bettina Zacharias (GESIS)
Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

# Content

methods, data, analyses | Vol. 14(2), 2020, pp. 159-194 DOI: 10.12758/mda.2020.05

# What Do You Think? Using Expert Opinion to Improve Predictions of Response Propensity Under a Bayesian Framework

*Stephanie Coffey*[1], *Brady T. West*[2], *James Wagner*[2] & *Michael R. Elliott*[2, 3]

[1] *Joint Program in Survey Methodology and U.S. Census Bureau*

[2] *Survey Research Center, Institute for Social Research, University of Michigan-Ann Arbor*

[3] *Department of Biostatistics, University of Michigan-Ann Arbor*

**Abstract**

Responsive survey designs introduce protocol changes to survey operations based on accumulating paradata. Case-level predictions, including response propensity, can be used to tailor data collection features in pursuit of cost or quality goals. Unfortunately, predictions based only on partial data from the current round of data collection can be biased, leading to ineffective tailoring. Bayesian approaches can provide protection against this bias. Prior beliefs, which are generated from data external to the current survey implementation, contribute information that may be lacking from the partial current data. Those priors are then updated with the accumulating paradata. The elicitation of the prior beliefs, then, is an important characteristic of these approaches. While historical data for the same or a similar survey may be the most natural source for generating priors, eliciting prior beliefs from experienced survey managers may be a reasonable choice for new surveys, or when historical data are not available. Here, we fielded a questionnaire to survey managers, asking about expected attempt-level response rates for different subgroups of cases, and developed prior distributions for attempt-level response propensity model coefficients based on the mean and standard error of their responses. Then, using respondent data from a real survey, we compared the predictions of response propensity when the expert knowledge is incorporated into a prior to those based on a standard method that considers accumulating paradata only, as well as a method that incorporates historical survey data.

*Keywords*: Bayesian Analysis, Response Propensity, Expert Opinion, Elicitation of Priors, Responsive Survey Design

Responsive Survey Design (RSD; Groves and Heeringa, 2006) relies on accumulating paradata (i.e. data about the process of collecting survey data, see Couper 2000, 2017) and response data in order to introduce changes to data collection protocols or tailor data collection features to specific cases. These changes are made in pursuit of a survey goal, such as quality improvement or cost control. Unfortunately, by relying only on the partial current data as it accumulates, predictions generated from this partial data may be biased (Wagner and Hubbard, 2014) and, as a result, decisions made based on these predictions can be inefficient or even harmful.

Recently, survey researchers have introduced Bayesian approaches (Schouten et al., 2018) to mitigate this bias by supplementing the current accumulating data with prior beliefs, generated from external data such as past implementations of the same survey or the survey methodological literature (West, Wagner, Coffey and Elliott, 2019). While priors generated from past implementations of the same survey may be the most informative for a particular survey, that solution is not always an option. New surveys, or surveys whose designs have changed dramatically, may need to develop priors from different data sources. West et al. (2019) explored using a literature review to source prior information for response propensity models in the National Survey of Family Growth (NSFG). While priors from the literature review did not perform as well as priors from historical NSFG data, they outperformed model predictions made only using current accumulating paradata, particularly in the middle portion of the data collection period.

The present study evaluates another potential source of prior information. Here, expert knowledge was elicited from survey managers ("experts"), through a self-response questionnaire designed to collect their predictions of attempt-level response rates, or changes in those expected response rates, for various types of sample members. Given those survey responses, pooled priors were created from expert respondent data. The structure of the items in the questionnaire completed

*Acknowledegments*

This work was supported by a grant from the National Institutes for Health (#1R01AG058599-01). The National Survey of Family Growth (NSFG) is conducted by the Centers for Disease Control and Prevention's (CDC's) National Center for Health Statistics (NCHS), under contract # 200-2010-33976 with University of Michigan's Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS nor the other funding agencies.

*Direct correspondence to*

Stephanie Coffey, Joint Program in Survey Methodology and U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746
Email: stephanie.coffey@census.gov

by the experts mimicked that of the existing response propensity model. We then evaluated these priors' ability to improve predictions of response propensity in the National Survey of Family Growth (NSFG) relative to only using partial data from the current round or using historical data as an alternative source for the development of priors. This manuscript discusses the content of the questionnaire, the identification of experts, the method for generating priors, and an evaluation of how the information from expert elicitation affects the bias and root mean squared error (RMSE) of the daily predictions of response propensity. We found that priors based on expert opinion led to modest improvements in prediction during the middle and late portions of data collection when compared to using only current round data. Additionally, we found that priors based on expert opinion were sometimes competitive with, though generally did not outperform, an approach that used historical data evaluated in West et al. (2019). We also identified several ways to improve upon our elicitation process that may lead to further improvements in predictions based on expert opinion over methods more commonly used in RSDs.

# Background

## Responsive Survey Design

Responsive survey design (RSD; Groves and Heeringa, 2006) has emerged as a framework for maintaining or improving survey outcomes in an increasingly difficult survey climate. Increasing data collection costs, and decreasing cooperation and response rates, have caused survey methodologists and managers to explore alternatives to the prevailing "one path fits all sample members" approach to data collection operations (Axinn, Link and Groves, 2011). Instead, RSD uses accumulating paradata and response data to make changes to later data collection protocols. These changes attempt to increase data quality in some specified way or control costs, relative to continuing with the standard data collection protocol. Types of protocol changes may include introducing another mode (Coffey, Reist and Miller, 2019), changing the effort spent on specific cases (Rosen et al., 2014), or a change in tokens of appreciation combined with subsampling (Wagner et al., 2012).

In an RSD, one of the most common ways to tailor data collection features to specific cases is with predicted propensity scores. Based on frame data and accumulated paradata, these predictions can be used to alter data collection operations. Various surveys have utilized propensity scores to differentially implement a variety of data collection features, including protocol assignment (Peytchev, Rosen, Riley, Murphy and Lindblad, 2010; Roberts, Vandenplas and Stahli, 2014), incentives (Chapman, 2014), and allocation to nonresponse follow-up (Laflamme and

Karaganis, 2010; Thompson and Kaputa, 2017) in hopes of improving survey outcomes.

Paradata from the current round of data collection provide useful predictors of survey outcomes, such as response propensity, for the sampled cases currently receiving recruitment effort. In an RSD, targeted interventions are applied to cases during the data collection period in order to shift response propensities in pursuit of a cost- or quality-related survey goal, necessitating high quality predictions of these propensities. However, during the survey period when an RSD would be implemented, the accumulating paradata are "incomplete" relative to the final data, in that completed cases and incoming data from early in the data collection period may not be representative of that which will be collected later in data collection. As a result, only using the accumulating data from the current round of data collection could result in biased predictions of response propensity (Wagner and Hubbard, 2014) or reduced prediction performance when predicted propensities are classified into response categories, either of which could lead to inefficient decisions. In this paper, we focus on the error in the predictions of response propensity scores, as opposed to the secondary step of classification error.

In order to improve predictions, survey practitioners often use external data that may be more representative of a full data collection period. It is relatively common to estimate the coefficients of a predictive model using historical data, such as a prior implementation of the survey, and then apply those coefficients to the current round of data collection (Schouten, Calinescu and Luiten 2013; Schouten, Wagner and Peytchev, 2017; Schouten, Mushkudiani, Shlomo, Durrant, Lundquist and Wagner, 2018). While this method provides data that might be representative of an entire data collection, it ignores current data in the prediction process.

More recently, survey researchers have begun exploring Bayesian approaches that utilize both external and current data in the prediction process. Prior beliefs are generated from external data, most commonly historical data from the same survey, and those priors are then updated as the current data accumulates. Schouten et al. (2018) discuss using Bayesian methods for predicting response and cost under different scenarios. Through simulation, they demonstrate value in the Bayesian methods in terms of reduced RMSE of predictions, while stressing that misspecification of the priors with respect to the true data should be relatively small. Empirical evidence is also emerging (West et al., 2019) that combining published estimates or historical information and current round information in a Bayesian setting can improve prediction.

## Empirical Evidence and Sources of Prior Information

West et al. (2019) compared the performance of predictions of response propensity in the NSFG, a nationally representative quarterly survey in the U.S., when Bayes-

ian methods are used versus when only current data is used. The Bayesian methods incorporated external information in the form of priors, either from past implementations of the NSFG or from published research on propensity models found through a literature review. Results demonstrated that the Bayesian approaches consistently reduced both the bias and the mean squared error (MSE) of predicted response propensities, particularly in the middle of data collection, when an RSD may be implemented. This was true for either source of prior information -- the historical data or the literature review.

The quality of the prior information is directly related to its ability to improve predictions of interest, and so the source of prior information is an important consideration. It seems reasonable that historical data from the same survey would result in the most informative priors for the prediction of interest; however, there may be cases where this information is not available. New surveys, for example, would not have access to historical information. Additionally, surveys that have undergone significant redesign, such as introducing a new mode, changing an incentive amount, or dropping a screening interview, may find that priors based on historical paradata are no longer available.

There may be cases where even a literature review produces limited or no useful external information. In the case where a survey has an unusual or unique target population, or the prediction of interest is not as common as response propensity, there may not be sufficient information in the literature from which to develop priors. In these cases, where there is an absence of objective information, expert opinion may be the only option for generating the necessary information for prior construction. Expert opinion is often used implicitly in survey planning – experienced survey managers may provide input into expected response rates to help determine sample sizes, or for estimating budgets. Additionally, they may help explain variation in progress or response rates during data collection. Transforming expert opinion into priors explicitly incorporates this information into the prediction model.

## Expert Elicitation

Clinical trials and health care evaluations often rely on prior beliefs for a variety of reasons. Dallow, Best and Montague (2018) describe a protocol for eliciting expert opinion in order to improve the drug development process. Mason et al. (2017) propose a practice for leveraging expert opinion in the analysis of randomized controlled trials when there are missing observations for patients. Additionally, Boulet et al. (2019) demonstrate the use of expert opinion in a variable selection process for personalized medicine. When novel treatments are tested, or prior trials have very small sample sizes or are otherwise not comparable, expert opinion can be relied upon for developing priors (Hampson, Whitehead, Eleftheriou and Brogan, 2014).

Spiegelhalter et al. (2004, Ch. 5) as well as O'Hagan (2019) provide overviews of the expert elicitation process, and the potential biases that may arise in priors elicited from individuals. *Availability bias* may arise when experts are asked about easily recalled events – they may estimate a higher or lower probability than is accurate. For example, if survey experts have recently seen frequent reports of language barriers along with increasing non-interview rates, the experts may inflate the effect that a language barrier has on overall response rate or response propensity, even if there are other contributing factors to increasing non-interview rates. *Anchoring bias* may lead experts to shrink intervals between different categories or groups based on a provided piece of information or their initial elicited quantity or probability. Once an expert learns from the elicitation instrument, or offers through the elicitation process, that the expected response rate for one group is 45%, future answers about different subgroups may be biased towards 45%.

*Overconfidence bias* may lead to distributions of the priors with insufficient variance. This may occur when elicitation happens in small groups and some strongly opinionated experts convince others of their opinion, a behavior also known as groupthink. Alternatively, in individual elicitation, overconfidence bias may arise because of the expectation of experts that they have, in fact, a greater amount of expertise than they actually do, resulting in under-reported uncertainty. *Conjunction fallacy bias* may arise when a particular event is given a higher estimated probability when it is the subset of another event. For example, on any given contact attempt, the probability that any open case will have had a callback request and respond is necessarily smaller than the probability that any open case will respond. However, an expert may suggest the opposite, thinking that having a callback request makes response much more likely. This bias is often due to the rarity of one of the two events, which in this case would be the callback request. Finally, *hindsight bias* may arise if the expert is asked to provide a prior expectation after looking at the current data. Awareness of all of these types of bias is useful in the design of the expert elicitation process.

Spiegelhalter et al. (2004, Ch. 5) also discuss four common methods for elicitation: informal discussion, structured interviewing, structured questionnaires, and computer-based elicitation. Each of these methods requires different amounts of interaction with experts, and allows for different levels of complexity of prior development. Additionally, these authors discuss three methods for combining information when multiple experts are utilized: arriving at a consensus value among all experts, arithmetic pooling, or retaining individual priors. O'Hagan (2019), whose elicitation method elicits distributions from experts, discusses the combination of those distributions to generate a pooled empirical distribution for the prior.

Here, we adapted the concept of expert elicitation of priors from the clinical trials literature. Our goal was to evaluate whether expert opinion can be helpful when little objective data is available for generating priors for the coefficients in a

logistic regression model used to estimate propensity of response. In this application, we elicited opinion from experts independently through an internet questionnaire, and used arithmetic pooling to combine the elicited information into priors for models used to generate daily predictions of response propensity in the NSFG.

# Data and Methods

## Overview of the National Survey of Family Growth

The NSFG is conducted by the National Center for Health Statistics, under contract with the Institute for Social Research (ISR) at the University of Michigan. The NSFG, in its current iteration, is a cross-sectional survey for which data were collected continuously throughout the calendar year from 2011-2019. In a given year, four data collection operations are conducted, with data being collected from four independent, nationally representative samples. The field operations for each sample last three months, or one quarter (e.g., January to March, April to June). The survey selects a national sample of U.S. housing unit addresses each quarter of the year. The target population from which the NSFG selects these four independent national samples is 15 – 49 year old persons living in the U.S. (Lepkowski, Mosher, Groves, West, Wagner and Gu, 2013). The NSFG is a two-stage survey, meaning there is first a screener interview to determine eligibility, followed by the main interview. Interviewers first visit randomly sampled households and attempt to screen the households for eligibility. Within eligible households, one of the eligible individuals is randomly selected to complete the main survey interview, which usually takes 60-80 minutes and covers a variety of fertility-related topics.

NSFG paradata are aggregated on a daily basis and used to predict the probability that active households will respond to either the screening interview or the main interview. Survey managers might use these predictions for prioritization of active cases (e.g., Wagner et al., 2012) or for stratifying the sample when selecting a subsample of active cases for the new data collection protocol after 10 weeks (Wagner et al., 2017). At this point, managers may oversample high-propensity cases, or offer a higher token of appreciation to encourage response. Accurate model-based predictions are thus essential for maximizing the efficiency of the data collection effort in any given quarter. For purposes of this study, we focus on models for the probability of responding to the initial screening interview.

## Response Propensity Models in the NSFG

For this application, we used data from five quarters of the NSFG (Quarters 16 – 20), covering the June 2015 to September 2016 time period. For each of the five

quarters, our prediction of interest was the probability of response to the screening interview at the next contact attempt, using either the current accumulating paradata only, or the combination of priors generated from expert elicitation and the current accumulating paradata. We also compared these methods to the best performing method in West et al. (2019), which combined current accumulating paradata with priors based on historical data from the eight preceding quarters of data collection.

In order to compare predictions generated from our proposed method with those discussed in West et al (2019), we used the same predictive modeling approach (discrete time logistic regression), and the same set of predictors of screener response propensity. In that paper, eight quarters (or two years) of the NSFG (Quarters 13 – 20) were combined into a stacked dataset containing all contact attempt records and a binary outcome for each record that indicated whether the screener interview was completed on that particular attempt or not. The authors then fit a discrete time-to-event logistic regression model to this dataset to identify significant predictors. Available predictors included sampling frame information, linked commercially-available data, and NSFG paradata, all of which have been used to predict response propensity in the NSFG (West, 2013; West and Groves, 2013; West et al., 2015). The authors used a backward selection approach to model-building, retaining all predictor variables that appeared in all eight quarters with a p-value less than 0.05 based on a Wald test for all regression parameters associated with a given variable.

They then included two predictor variables that were important for sampling and weighting in order to control for sampling domain in the response propensity model. The first was the sociodemographic domain of each housing unit, based on the percentage of the population in the Census Block Group containing the segment that is Black and/or Hispanic as reported in U.S. Census data. The second was a three-level categorical variable indicating whether a case was in a self-representing area, a non-self-representing metropolitan statistical area (MSA), or a non-MSA non-self-representing area. Self-representing sampling areas are geographic sampling domains that are large enough to be sampled with certainty in a probability proportionate-to-size sample, and, therefore, represent only themselves during weighting and estimation. These two variables were initially included in the backwards selection procedure, but were not found to be statistically significant, and so were not retained. However, after consultation with data collection managers, these two variables were added back into the response propensity model in order to control for sampling domain in the predictive model.

All retained predictors from the backward selection process carried out in West et al. (2019), including their estimated coefficients and standard errors, are listed in table A1 in the online appendix. Several predictors came from each available data source: the sampling frame, commercially-available data, and paradata.

By using the same list of predictors, and the same discrete-time logistic regression model specification, we are able to compare the effect that priors based on expert elicitation have on the predictions of response propensity, versus excluding prior information, or using priors from historical data. The focus of our analysis is on the relative performance of these methods given a particular model.

## Design of Prior Elicitation Process

For this proof-of-concept study, we wanted our prior information to be based upon a relatively large group of experts to generate a reasonable distribution from which to estimate priors. Our target sample size meant that elicitation methods requiring significant interaction with experts, including informal discussion and structured interviewing, were not feasible. As a result, we created and distributed a structured questionnaire to selected experts, who could then respond at their convenience. The questionnaire asked experts to provide their opinions on attempt-level response rates for subgroups with various types of characteristics, and, in some cases, opinions on changes to response rates based on certain characteristics.

The questionnaire included the significant predictors found in the retrospective analysis of the NSFG response propensity model, as described in Section 3.2. These predictors include items from the sampling frame, including geographic and sampling strata information, as well as time-varying attempt-level information, derived from accumulating paradata. Fixed characteristics include sampling frame or commercially available data, like the 9-level Census Division geographic variable. In the questionnaire, we asked experts their opinions on their expected response rates for each of the nine categories. Time-varying covariates were based on paradata and include indicators for past contact or instances of the sample member expressing questions, comments or concerns. In the questionnaire, we requested information about the expected change in response rate for characteristics like each additional contact attempt, or whether the sample member expressed comments on concerns on the most recent contact attempt. We also asked experts to provide their experience with survey data collection by selecting one of three categories: 0 to 4 years, 5 to 15 years, and 15 or more years.

We solicited feedback from two survey experts prior to distributing the questionnaire in order to get basic feedback about content, complexity, and readability. In some cases, edits resulting from this initial feedback changed the format of the questions to make them easier to understand and answer. This meant that the format of the questions did not always match the format of the predictor in the propensity model. The final version of the questionnaire can be found in the online appendix, and in the Center for Open Science repository (https://osf.io/3kxzb) at the Open Science Framework (log-in required).

Given the target number of experts, we opted to develop priors through arithmetic pooling of all respondent information. At the same time, we wanted to avoid the biases mentioned by Spiegelhalter et al. (2004, Ch. 5). In order to avoid anchoring bias while still eliciting reasonable responses, we provided an overall expected attempt-level response rate (24%), but did not provide anchor points for any particular category in the survey, allowing the experts to provide input for all items and categories. To avoid hindsight bias (Schouten et al., 2018) arising from the fact that experts at ISR also conduct the NSFG, we recruited additional experts from the U.S. Census Bureau (Census). These additional experts have experience managing interviewer-administered data collections, but do not have experience with the NSFG or its data. By soliciting predictions from two geographically dispersed survey organizations with varying familiarity with the NSFG, we also hoped to protect against overconfidence bias (Schouten et al., 2018), which can lead to prior distributions that are too narrow and do not accurately reflect the uncertainty in the prior.

At both ISR and Census, we worked with senior survey managers to identify experienced interviewer supervisors, field directors, and survey methodologists who were knowledgeable about survey processes and reviewed progress data on a daily basis as part of their job responsibilities. We recruited eight individuals from ISR, and 12 from Census (two from each of the six regional offices). During March 2019, the recruited experts were asked to complete the questionnaire, and were encouraged to provide feedback, either directly or through a scheduled debriefing. We summarize the feedback received in the Results section.

## Method for Deriving Priors

We obtained 20 sets of expert responses about the effects on attempt-level response rates of various characteristics of sample members and paradata items, subject to some item nonresponse. We used arithmetic pooling to combine the priors and generate an expected mean and standard error for a coefficient in an attempt-level response propensity model (Spiegelhalter et al., 2004, Ch. 5).

Before pooling, however, we had to convert the estimates of differences in response rates to model coefficients for use in a logistic regression model. When categorical variables are included as predictors in a logistic regression model, the estimated coefficients are generally interpreted with respect to a reference category. Therefore, the mathematical manipulation involved identifying a reference category, calculating odds ratios with respect to the reference category, and then taking the natural log of the odds ratio to obtain a logistic regression model coefficient, or beta. We first did this for each respondent's information individually.

Formula 1 below demonstrates how to calculate the coefficient for the $k^{th}$ category of the $j^{th}$ item for the $i^{th}$ expert, $\hat{\beta}_{ijk}$, given the estimated probability of

response for category $k$ of interest, $\hat{p}_{ijk}$, and the estimated probability of response for a reference category $R$, $\hat{p}_{ijR}$.

$$\hat{\beta}_{ijk} = ln\left(\frac{\hat{p}_{ijk} / \left(1-\hat{p}_{ijk}\right)}{\hat{p}_{ijR} / \left(1-\hat{p}_{ijR}\right)}\right) \tag{1}$$

Using gender as an example (abbreviated $G$ in the expression below), assume that the $i^{th}$ respondent estimates the expected call-level response rate for female sample members to be 85% (as opposed to 70% for males), and male is the reference category. The *beta* for female sample members, for the $i^{th}$ expert, would be:

$$\hat{\beta}_{iGF} = ln\left(\frac{\hat{p}_{iGF} / \left(1-\hat{p}_{iGF}\right)}{\hat{p}_{iGM} / \left(1-\hat{p}_{iGM}\right)}\right) = ln\left(\frac{0.85 / \left(1-0.85\right)}{0.70 / \left(1-0.70\right)}\right) = 0.8873$$

Continuous variables were converted to model parameters using the same formula but with a slightly different explanation. For these items in the questionnaire, expert opinion was elicited about the *change* in response propensity, given some unit change in the continuous variable. For example, survey managers were asked to provide their expected change in response rate for each additional contact attempt made on a sample member, and a survey manager might have responded saying they would expect a -10% change, or a 10% reduction, in response propensity for each additional contact attempt.

However, unlike standard linear regression, where there is linear change for every unit increase, logistic regression results in exponential change for each unit increase, meaning the change in response propensity is dependent on *which* unit increase is being considered (e.g. from 1 to 2 attempts, or from 8 to 9 attempts). In the case of continuous variables, we did not have a defined reference category, and so the reference is always to the average attempt-level response rate of 24%.

If the $i^{th}$ expert believes that increasing the number of contact attempts, $j$, by one would change the attempt-level response rate by some amount, we can adapt Equation (1) above for a continuous variable. While we do not have a defined reference category, we have the overall average attempt-level response rate, 24% and the expected change provided by the expert, 5%. This results in a model coefficient of:

$$\hat{\beta}_{ij} = \ln\left(\frac{odds\left(attempts = \left(n+1\right)\right)}{odds\left(attempts = \left(n\right)\right)}\right) = ln\left(\frac{0.29 / 0.71}{0.24 / 0.76}\right) = 0.2573 \, .$$

We note at this point that, while we have elicited priors on a linear scale, linking these back to the logistic scale changes the interpretation. We provide more consideration of this issue in the Discussion section.

To pool the expert information, we then took an arithmetic mean, $\hat{\bar{\beta}}_{jk}$ (or $\hat{\bar{\beta}}_{j}$ for continuous items), of the coefficients from the expert respondents. The standard error of the prior, $SE\left(\hat{\bar{\beta}}_{jk}\right)$, was estimated by dividing the standard deviation of the coefficients from the respondents by the square root of the number of respondents, *n*.

$$\hat{\bar{\beta}}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_{ijk} \tag{2}$$

$$SE\left(\hat{\bar{\beta}}_{jk}\right) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\beta}_{ijk} - \hat{\bar{\beta}}_{jk}\right)^2} \tag{3}$$

We chose to transform each expert response into an odds ratio, take the log, and then pool the individual log-odds ratios for a few reasons. Mathematically, by first transforming each expert response into a log-odds ratio before pooling, we are working under the assumption that the log-odds are normally distributed, as opposed to the response rate or response propensity, which is how the experts provided their opinions. We felt this assumption was reasonable. First, response rates and response propensities are bounded at (0,1), and are not normally distributed, whereas the log-odds can take on any number on the real line. Additionally, the log-odds is a linear function, while the function for the odds (and for probabilities) are multiplicative and exponential, which suggests that the log-odds might converge to a normal distribution more quickly than the odds, given enough sample size.

Operationally, by generating a model coefficient for each expert, we were able to calculate a mean and standard error for each model coefficient. If we had first taken the mean of the expert response first, and then transformed that estimate to obtain our model coefficient, we would no longer be able to generate a variance, as we would have only one estimate.

For each covariate of interest, we used $\left(\hat{\bar{\beta}}_{jk}, SE\left(\hat{\bar{\beta}}_{jk}\right)\right)$ to define a normal prior distribution in our prediction models. Each prior was based on a maximum of 20 responses, but item-level nonresponse reduced the number of responses to varying degrees (see Table A2 for individual response counts). Due to the small sample sizes, we ignored the potential covariance between the coefficients, resulting in a variance-covariance matrix that is only non-zero on the diagonal. This is different from the methods evaluated in West et al. (2019) that utilize historical data to generate priors. For those methods, including the historical method replicated in our results, estimated covariances were generated from the existing historical data.

Table A2 in the online appendix provides the prior information, $\left(\hat{\bar{\beta}}_{jk}, SE\left(\hat{\bar{\beta}}_{jk}\right)\right)$, for each covariate included in the propensity models, provided that there were at least three contributing respondents. Further, an Excel spreadsheet available in the online supplementary material provides a template for estimating these priors for

the survey items in the propensity model. For demonstration purposes, simulated data are included in the table, including missing cells, which would occur should an expert not respond to a particular question.

## Methods for Predicting and Evaluating Response Propensities

Each of the five NSFG quarters of interest (Quarters 16 through 20, representing June 2015 - September 2016) were analyzed independently to introduce replication in our analysis. First, we used the expert opinions to generate the prior distributions for the response propensity model coefficients as described above. These priors were used for all five quarters.

We generated our "target" prediction at the case level for each of the five evaluation quarters by fitting a discrete time-to-event logistic regression model using the predictors identified in the backward selection model discussed in Section 3.2 to all contact attempt records from that quarter. This allowed us to estimate a "final" probability of responding to the screener interview at the last contact attempt for each case. Because this model uses all available information for a given quarter, we consider this the benchmark against which the prediction methods under evaluation will be compared. Table 1 below shows the ROC-AUC values when all contact attempt records were used to predict final response.

These model fit statistics reflect the in-sample performance of the models and demonstrate that the variable selection procedure from West et al. (2019), where these statistics are extracted from, yielded a reasonable list of predictors for our target response propensity. From that point, we are concerned with the case-level differences from the target propensity that the different methods produce.

Then, we generated daily predictions of response propensity based on contact history data accumulated prior to each day. Our baseline predictions came from the model using only accumulating current round paradata. Our proposed predictions came from the model that also incorporated prior information from expert opinion. Additionally, we included predictions that incorporate prior information from historical data, as presented in West et al. (2019). In that paper, the authors found that

*Table 1* Model Fit Statistics for In-Sample Predictions of Response, 5 Evaluation Quarters

| | Q16 | Q17 | Q18 | Q19 | Q20 |
|---|---|---|---|---|---|
| ROC-AUC | 0.711 | 0.682 | 0.661 | 0.690 | 0.654 |
| Nagelkerke-Pseudo $R^2$ | 0.143 | 0.115 | 0.089 | 0.130 | 0.086 |

the historical data method performed the best in their application. We include the historical data method here so we can understand how well the expert elicitation method performs when compared to both the "current data only" method and one of the historical data methods evaluated in West et al. (2019).

Prediction of daily response propensity for each of these three methods is carried out just as it would have been if the approach were to be employed during data collection. For each of the five quarters of interest, we use the accumulated contact attempt record information (with a screener response indicator for each record) up to day *d* to estimate the coefficients for the discrete time logistic regression model for that data collection period. Then we use those coefficients to predict the response propensity at the next contact attempt for all cases who were nonrespondents on day *d*. We repeat this for each day of data collection from Day 7 to Day 84.

Using only the current quarter of paradata, the response propensity, $\hat{p}_{id}$, was modeled as follows:

$$\hat{p}_{id} = \hat{p}\left(y_{id} = 1 | X_{id}\right) = \frac{\exp\left(\Sigma_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}{1 + \exp\left(\Sigma_{v=0}^{V} \hat{\beta}_v X_{idv}\right)} \tag{4}$$

where $y_{id}$ is the response status for the $i^{th}$ case after a contact attempt on the $d^{th}$ day, and $X_{id}$ is the set of predictors $v$ for the $i^{th}$ case after the $d^{th}$ day. These predictors may be fixed (e.g., geographic predictors) or time-varying (e.g., prior contact status). The $\hat{\beta}_v$ are estimated coefficients for the $X_{idv}$ predictors. They are estimated from the likelihood in equation (5) based on the contact attempt records that have been accumulated through day *d*.

$$L\left(\hat{\beta}_0, \ldots, \hat{\beta}_v\right) = \prod_{i=1}^{n} \prod_{j=1}^{d} \left( \frac{\exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}{1 + \exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)} \right)^{y_{id}} \left( 1 - \left( \frac{\exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}{1 + \exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)} \right) \right)^{(1-y_{id})} \tag{5}$$

The only difference between the target prediction and the baseline, current-data only method is the time at which the prediction is made. For the target predictions, all contact attempt records from a given quarter are used (*d* is after the last contact attempt is made in a given quarter); for the baseline method, only data accumulated through day *d* are used.

In a Bayesian setting (Gelman et al. 2013), the likelihood matches the frequentist formulation. The only estimated parameters in this expression are the $\hat{\beta}_v$, and so these are the parameters for which priors are defined. As described in Section 3.4, we assumed a normal distribution, $\beta_v \sim N\left(\mu_v, \sigma_v^2\right)$, for our priors with the mean and variance based on our expert elicitation procedure. The posterior multi-

plies the prior over the parameters in the likelihood to combine the information, as shown in equation (6):

$$pos\left(\hat{\beta}_0,\ldots,\hat{\beta}_v\right)=\prod_{i=1}^{n}\prod_{j=1}^{d}\left[\left(\frac{\exp\left(\sum_{v=0}^{V}\hat{\beta}_v X_{idv}\right)}{1+\exp\left(\sum_{v=0}^{V}\hat{\beta}_v X_{idv}\right)}\right)^{y_{id}}\left(1-\left(\frac{\exp\left(\sum_{v=0}^{V}\hat{\beta}_v X_{idv}\right)}{1+\exp\left(\sum_{v=0}^{V}\hat{\beta}_v X_{idv}\right)}\right)\right)^{(1-y_{id})}\right] \quad (6)$$
$$\times\prod_{v=0}^{v}\frac{1}{\sqrt{2\pi\sigma_v^2}}exp\left(-\frac{1}{2}\left(\frac{\beta_v-\mu_v}{\sigma_v}\right)^2\right)$$

In the Bayesian version of the prediction, it is clear that the priors add additional information to the prediction. This can be beneficial when the likelihood is based on very sparse data, or partial data that are not representative of the full data collection process, both of which occur earlier in the data collection process. Code in the SAS 9.4 programming language that can be used to carry out these predictions is available in the online supplementary materials.

For each method, we will compare predictions for each contact attempt on each day of the data collection quarter to the "target" predictions (based on all cumulative data) in order to generate daily estimates of the bias and root mean squared error (RMSE) for the predictions. The mean daily bias for the $m^{th}$ method is defined as:

$$B^m=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\rho}_i^m-\rho_i\right) \quad (7)$$

and the daily RMSE for the $m^{th}$ method is defined as:

$$RMSE^m=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\rho}_i^m-\rho_i\right)^2} \quad (8)$$

We then summarized those estimates using boxplots for three different parts of data collection: early (day 7 – 30), middle (day 31 – 60), and late (day 61 – 84).

The end-of-data-collection response propensity is not the only possible target, but this choice does allow us to evaluate whether the use of Bayesian approaches with informative priors can reduce error in the predictions of response propensity at a given contact attempt versus using only current round paradata. Additionally, we will be able to evaluate whether the use of expert opinion (in the absence of historical data) can perform similarly to the historical data, were it available.

# Results

## Descriptive Statistics for Selected Priors

We first wanted to understand if ISR experts have different expectations than Census experts, potentially due to the varying familiarity with NSFG or simply being a part of a different survey organization. We also collected information about the experts' length of experience with survey data collection, thinking opinion may vary with length of experience and more experienced managers may provide more useful information. We then examined distributions of the individual experts' betas, generated using Equations (1) and (2) above, by organization and experience level. Here we provide examples of these distributions to illustrate similarities and differences in the provided opinions. Due to the small sample sizes, we do not provide tests of significance with respect to these differences. Instead, we are interested in the means and general trends of the expert opinion by category in order to understand, at a high level, if different types of experts provide different information.

We first examined distributions of coefficients related to two time-varying covariates, Contact Status and Concerns Status. Contact Status had three possible response categories: if there was ever contact with the sample member, contact on the previous attempt, or if there had never been contact with the respondent, which was used as the reference category. Concerns Status had four possible response categories: if concerns were ever expressed by the sample member, if concerns were expressed on the previous visit, if strong concerns were ever expressed, or if no concerns were ever expressed (the reference category). We looked at how responses differed by organization (Figures 1 and 3) and level of experience (Figures 2 and 4).

For both variables, we found largely the same results. There were no large differences found in the point estimate for the priors by survey organization, shown in Figures 1 and 3.

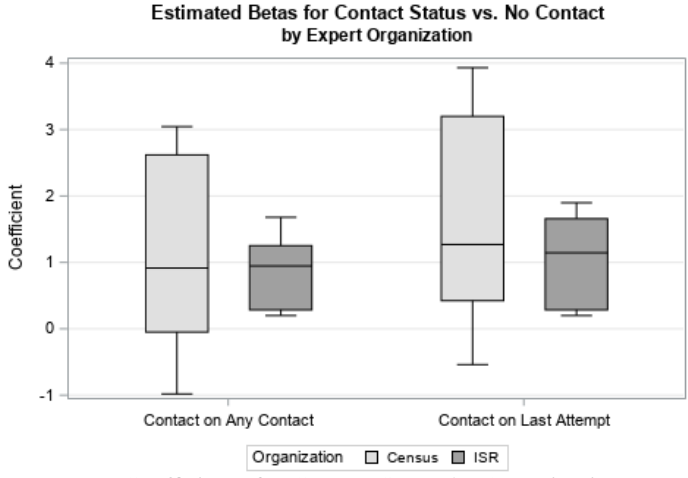


*Figure 1* Coefficients for Contact Status by Organization

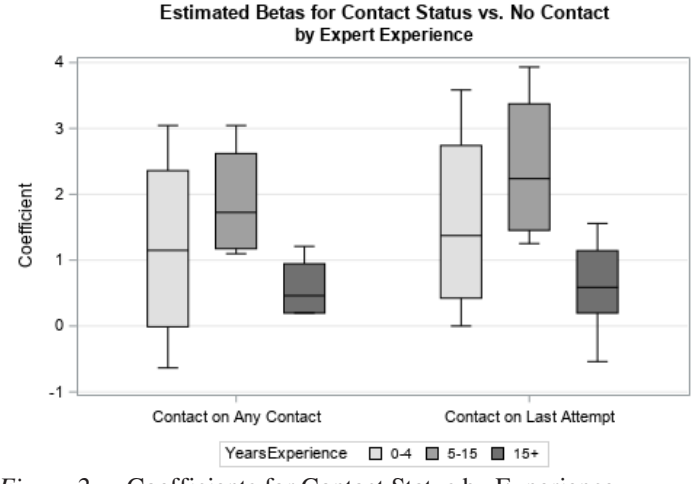


*Figure 2* Coefficients for Contact Status by Experience

When examining the priors by level of experience (Figures 2 and 4), interviewers with 0-4 or 5-10 years of experience generated similar point estimates for the betas, while experts with fifteen or more years of experience showed differences with respect to the point estimates. Specifically, experts with 15 or more years of experience appear to perceive, on average, that any one covariate has less of an impact on response propensity than do experts with less experience.

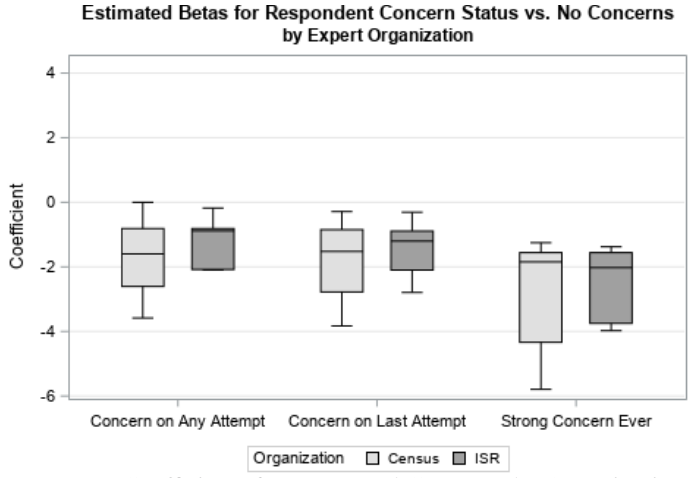


*Figure 3* Coefficients for Expressed Concerns by Organization

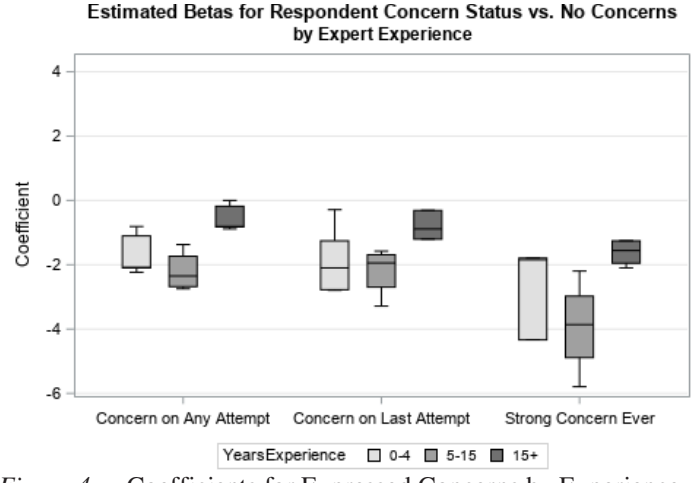


*Figure 4* Coefficients for Expressed Concerns by Experience

Other questionnaire items showed more clear differences between the survey organizations. Figure 5 shows the effect of various types of listing procedures on response propensity, versus listing alone on foot. Here, there are not only differences in the means by survey organization, particularly for listing in a car with another person and on foot with another person, but the means are in the opposite directions from the reference category, and the Census Bureau estimates are highly variable compared to estimates from ISR. In this particular case, feedback showed that Census Bureau experts did not see a link between listing method and response

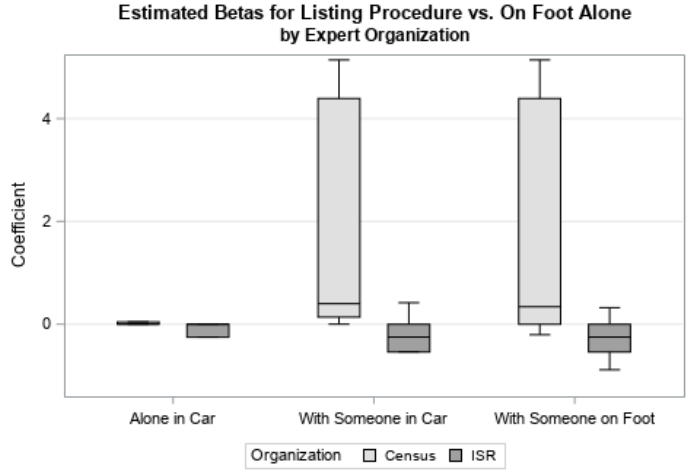


*Figure 5* Estimated Betas for Listing Procedure by Organization

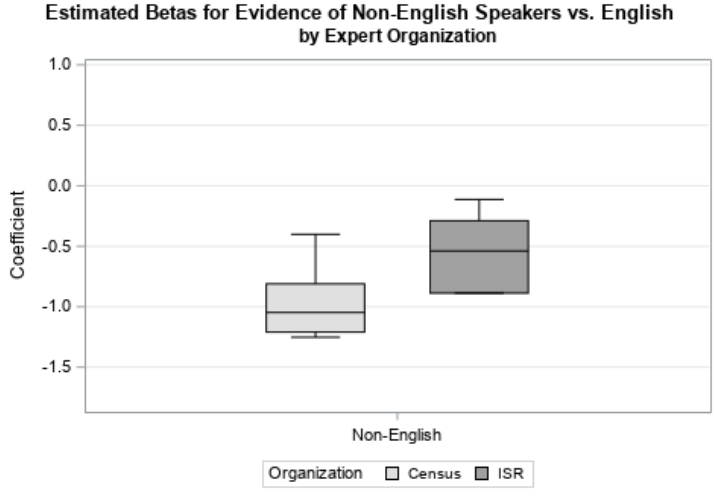


*Figure 6* Estimated Betas for Likely Non-English Speaker by Organization

propensity, resulting in highly variable responses. We discuss the additional expert feedback that we received on the survey more in Section 5.

Figure 6 displays the distributions of the betas by survey organization for the effect of evidence of a language other than English being spoken at home. Here, Census Bureau experts feel that evidence has a more negative effect on response propensity than ISR experts do. This may have to do with differences in the availability of bilingual interviewers or language specialists.

Understanding these similarities and differences is important for selecting the most appropriate experts to interview. Depending on the survey of interest, it

might be more important to select interviewers with specific skill sets, such as language specialties. It may also affect which questions are included on the questionnaire, or which priors are actually used in the prediction model. In the case of listing procedure, the feedback obtained might suggest ignoring the prior information for some or all of the experts, and either using an uninformative prior or dropping the variable from the model.

## Comparison of Methods

For each quarter, we treated the final prediction of response propensity, based on all accumulated contact data for the quarter, as the unbiased "target" prediction of response propensity. For each method, we then generate daily estimates of bias and RMSE with respect to the target prediction. Figures 7 to 12 display the performance of the Bayesian method using expert elicitation (EXPERT) to the current data-only method (Standard) and the precision-weighted prior Bayesian method (PWP) from West et al. (2019) that incorporates historical data. Our primary interest was to evaluate whether predictions generated using priors derived from expert opinion would be of higher quality than those generated using current data only, assuming historical data were not available for use. However, we were also interested in how the priors from expert opinion perform versus priors from historical data, which were evaluated in West et al. (2019). Because this was a retrospective analysis, we were able to examine both of these questions. Figures 7, 9 and 11 present the summarized distributions of estimated bias, while Figures 8, 10, and 12 present the summarized distributions of estimated RMSE.

Figures 7 and 8 focus on the early portion of data collection, from day 7 through day 30 (24 days). For each quarter, the 24 daily estimates of bias (Figure 7) or RMSE (Figure 8) were summarized using box plots. Early in data collection, the expert elicitation (EXPERT) method has a small but inconsistent effect on the bias and RMSE versus the standard method. For example, in quarters 19 and 20, the EXPERT method results in mean, median, and intraquartile ranges of both the bias and RMSE of the predictions that are slightly closer to zero than the Standard method, signifying an improvement. However, in quarter 16, the EXPERT method performs worse than the Standard method with respect to the mean and median values of bias and RMSE, and delivers no improvement in quarter 17. Overall, however, neither the PWP nor the EXPERT method offer consistent improvement over the Standard method early in data collection.

Figures 9 and 10 below represent the middle portion of data collection from day 31 to day 60. Beginning on day 31, there are noticeable reductions in the bias and RMSE of predictions for the EXPERT method. In all five quarters, the central tendencies of both the bias and the RMSE, as well as the intraquartile range, are shifted towards zero versus the Standard method. Further, in quarter 19, nei-

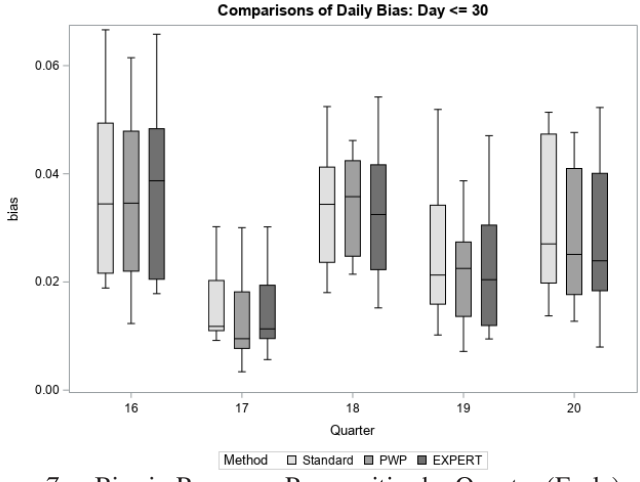


*Figure 7* Bias in Response Propensities by Quarter (Early)

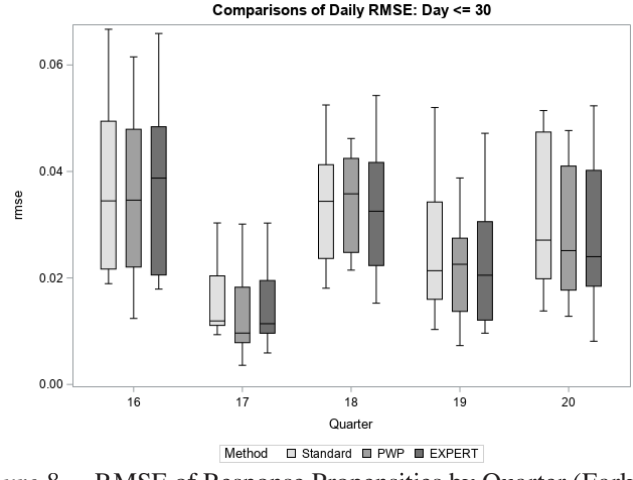


*Figure 8* RMSE of Response Propensities by Quarter (Early)

ther of the metrics have interquartile ranges that overlap between the Standard and EXPERT methods. For the most part, the PWP method continues to perform at least as well as the EXPERT method on measures of bias and RMSE, though the EXPERT method is certainly competitive, particularly in quarters 18 and 20. Here, unlike in the early portion of data collection, there is a clear benefit to using priors from expert elicitation if historical data are not available.

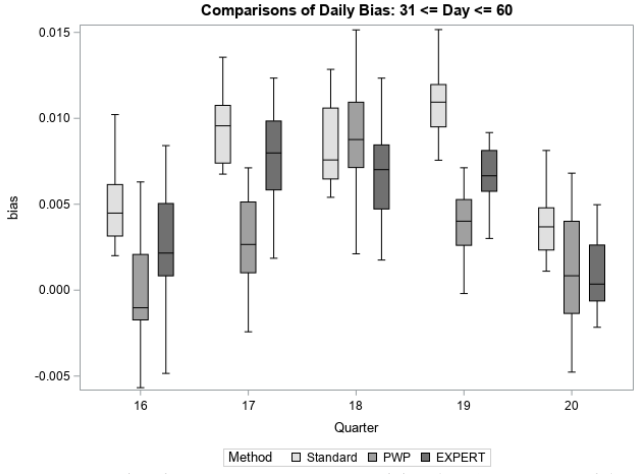


*Figure 9* Bias in Response Propensities by Quarter (Mid)

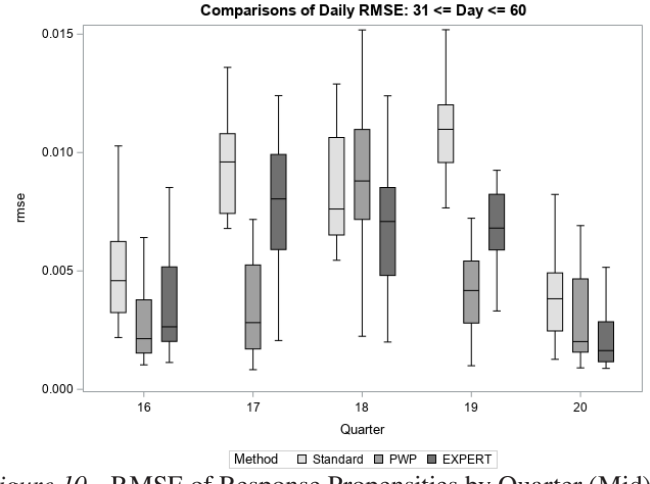


*Figure 10* RMSE of Response Propensities by Quarter (Mid)

During the final third of data collection, shown below in Figures 11 and 12, we continue to see that the EXPERT method leads to reduced measures of bias and RMSE versus the Standard method. These improvements are generally smaller than those found in Figures 9 and 10. Over the course of data collection, as more data are accumulated, it is likely that the Standard method improves in its ability to predict response, leading to smaller differences between the Bayesian methods and the Standard method. Additionally, it is more mixed as to whether the historical method or the expert opinion method is superior.

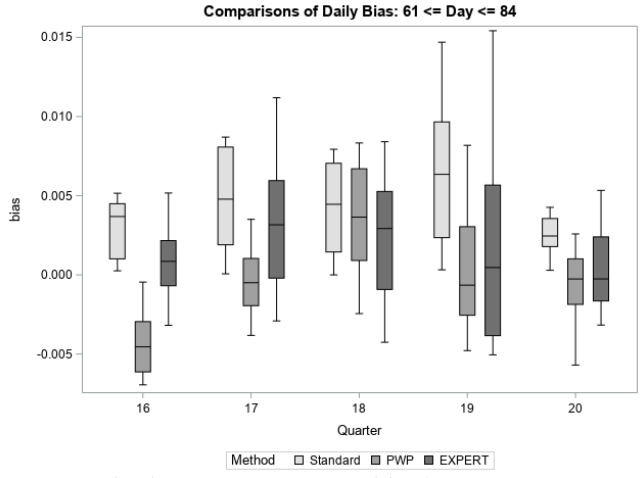


*Figure 11* Bias in Response Propensities by Quarter (Late)

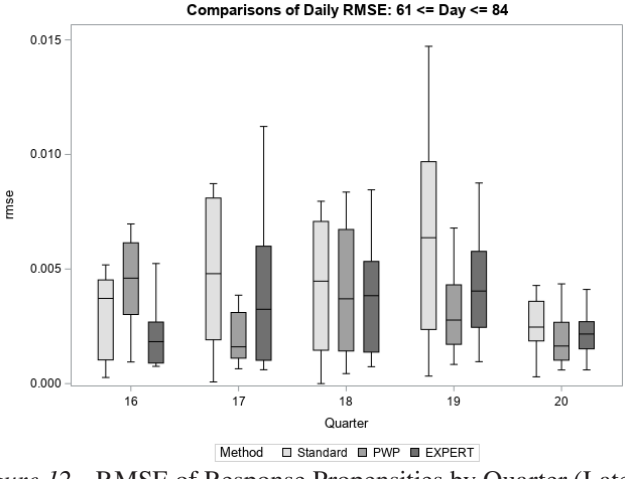


*Figure 12* RMSE of Response Propensities by Quarter (Late)

These results show that for this application, the PWP method results in the most consistent improvements in bias and RMSE of predictions of response propensity. However, the results also show that, in the absence of historical information, predictions that incorporate expert opinion still generally outperform the standard method, and can be a useful way to improve predictions of response propensity during data collection for the purposes of an RSD.

## Feedback from Survey Experts on Prior Questionnaire Development

Within two weeks of receiving questionnaire responses, we elicited feedback from experts in order to uncover issues with the questionnaire and identify potential areas for improvement. The experts had feedback in three main areas: the concepts identified in the questionnaire, how those concepts were translated into variables and categorical subgroups, and the lack of anchor points throughout the questionnaire.

The design of the questionnaire was driven by the variables available from the frame or from paradata. However, the concepts measured in the questionnaire did not always match concepts considered by the recruited experts. In our questionnaire, the experts provided two examples of this issue. In one instance, the predictive covariates from existing data sources were not meaningful concepts for survey managers. Mail Delivery Point Type is a categorical variable providing information on how mail is delivered to an address. This variable comes from the commercially available data and has several different categories that were significant in the variable selection model discussed in Section 3.2. However, when we included this variable (and all significant categories) on the expert questionnaire, only three out of 20 survey managers responded for any of the categories. During debriefing, survey managers explained that they did not have any experiential evidence that there was a relationship between response propensity and mail delivery. As a result, the survey managers generally declined to provide information for this concept.

On the other hand, survey managers explained that they do make use of concepts that were not included on the questionnaire. When providing feedback, one survey manager from the Census Bureau mentioned “perceived safety in a neighborhood” as a predictor of response propensity. In this case, this category was not included on the questionnaire because it was not a significant predictor in the response propensity model described in Section 3.2. It may be worthwhile to elicit information about predictors suggested by field experts, in order to capture information about predictors the experts find informative or predictive. This would allow confirmation that those particular items do not offer more explanatory power than the items retained from the propensity model.

In addition to defining meaningful concepts, it was also important to translate each concept into a variable that generated informative predictions, to the extent possible. This included determining whether a variable should be categorical or continuous, and, if categorical, how to define subgroups. Again, we found two clear examples of this issue. First, there were some instances where the categories that we provided in the expert questionnaire were not the same as those in the baseline model. As an example, age of householder, sourced from the sampling frame, was defined in the current model as having four categories: 18 - 44; 45 - 59; 60+; and

Missing. In the questionnaire, we only included three categories to simplify the response options: Under 50; 50+; and Missing. Age of the householder is provided on the sampling frame as a continuous variable, so in this instance, the different classifications posed no issues for generating predictions of response propensity. However, if the questionnaire included categories that were not able to be derived from the existing frame or paradata, the priors derived from expert information would not easily translate to covariates in the existing data.

The survey experts also suggested that the functional form of some of our variables was not ideal. For example, on the questionnaire, we asked the experts to predict the change in attempt-level response rates for every $10,000 increase in household income over the median. At least one expert suggested that the relationship was likely not linear, and a better way to elicit opinion might be categorical, such as using quartiles of household income. This would better represent what the experts suggested, which was that the top and bottom quartiles of household income would have a lower attempt-level response rate than those in the middle two quartiles.

The experts also provided feedback regarding anchor points. In designing the questionnaire, we made a conscious decision to only include the overall attempt-level response rate, 24%, in the introduction, leaving it up to respondents to generate all subgroup level response rates. This was primarily to avoid generating anchoring bias among the survey expert responses. However, while survey managers were comfortable ordering different subgroups of a variable, from highest to lowest predicted response rates, and even defining relative differences, they were less comfortable defining an initial response rate for one category, in order to then provide response rates that reflected the subgroup ordering and relative differences. We found evidence of this in the response data itself. Survey managers provided responses for nearly all questions, but on occasion, the predicted response rate ranges varied significantly (e.g., one manager might have all subgroup response rates in a range of 20% to 40%, while another would provide responses in a range of 60% or 80%). One survey manager suggested providing an anchor point for one subgroup in the categorical variable, from which they could then provide the relative differences for the remainder of the subgroups. We provided an overall anchoring point in order to facilitate estimates of effect levels. The 24% value acts as an "intercept" attempt-level response rate, from which specific categories of the questionnaire deviate. However, we did not provide any category-level anchor points in an effort to avoid anchoring bias. There was a concern that if we provided the overall attempt level response rate (24%) in addition to an anchor point for one of the categories, the experts would focus on the relationships between categorical response rates and the overall response rates. For example, had we provided the 24% overall attempt-level response rate, and a response rate of 35% for female respondents, the expert may ignore their own expertise to provide a response rate

around 13% in order to have the categorical response rates roughly match the overall attempt-level response rate. Our goal was to provide the minimum necessary amount of background information to allow the experts to use their own judgement to the fullest extent possible.

## Discussion

We hypothesized that in the absence of historical survey data, survey researchers would be able to generate priors from the experiences of survey managers that lead to improved predictions of response propensity over those made from just the data available for the current round of data collection. The results of this study demonstrate that eliciting expert opinion is a useful way to generate priors and improve prediction of response propensities. Particularly after the first month of the NSFG data collection process, priors generated from expert opinion resulted in predictions of next-contact response propensity with both lower bias and RMSE than predictions based on only current round data. One potential explanation for why the Bayesian methods did not improve the predictions in the first month of data collection is that the early experience in any quarter is highly variable. That is, in Bayesian terms, the likelihood varies from quarter to quarter in the first few weeks. The observed data are somewhat more stable after 30 days, but do not normally align with the final model until near 60 days into the quarter. Hence, it is during that interval – i.e. after the first 30 days but before the 60th day of the quarter – that the prior information is most useful.

This prior elicitation process is significantly more involved than building models from existing historical data. Developing a questionnaire, conducting data collection with survey experts, aggregating and organizing the response data, and generating priors may be time consuming, particularly as the number of covariates increases. As a result, eliciting expert opinion for generating priors may not always be the ideal solution. In our experience, the large majority of the time and effort was spent on the initial development of the questionnaire. We would expect changes, adaptations, and future implementations to require much less effort. Experts themselves spent, on average, less than an hour on the actual survey. Assuming a pay rate of $50 per hour, the actual elicitation portion of the survey would cost roughly $1,000. We can imagine numerous applications where this type of expenditure would be worth this cost, as in the case where a new survey has a specific target population that may not have coefficients well-estimated by the published literature. Further, this method may be useful for mathematically incorporating expert opinion into predictions of response rates for budgetary purposes, sample sizes, and power calculations. Given the high costs of face-to-face data collection, improved response propensity predictions may help data collection managers make better

decisions in an adaptive or responsive design framework. Evaluating of the ability of predictions based on such an approach to improve data collection outcomes is an interesting direction for future research. We are currently pursuing experimental work in this area.

Through the process of designing and implementing the questionnaire, debriefing the survey managers, and analyzing the collected data, we identified four areas survey researchers should consider when developing and implementing expert elicitation surveys. These areas include the selection of concepts for inclusion into the survey; the translation of those concepts into covariates and/or categories; the potential need for anchor points for categorical covariates; and lastly, the selection of experts for the survey. Attention to these areas will lead to information from experts that is more helpful for generating priors, which are ultimately combined with current data to generate posterior predictions of response propensity.

For this particular questionnaire, through debriefings and response analysis, we observed several opportunities for improvement in the design process for expert surveys. Mindful selection of concepts and the subsequent translation of categorical variables will help experts provide more informative prior expectations. By working with experts to determine which data fields on the frame and in the paradata effectively translate to concepts used by survey managers, the value of the elicited information may increase. Additionally, it may uncover concepts used by survey managers when developing ad hoc expectations for response propensities that are not currently provided by data systems. There may be an opportunity then for expert opinion to motivate a modification of existing systems, either by appending an additional piece of information from the survey frame (if available), or capturing this concept in paradata, potentially through interviewer observations.

In order for experts to provide opinions on attempt level response rates for a survey, particularly when they are unfamiliar with the exact topic questionnaire, it may be helpful to provide context to the survey managers about general attempt-level response rates, or even provide an anchor point for one category of a variable. Providing an anchor point for a particular subgroup may be a reasonable solution to this issue, but it may increase anchoring bias in the remainder of the experts' responses. Additionally, in the case of categorical covariates in a logistic regression, it may not be absolutely critical. Generating priors requires constructing odds ratios, using one subgroup as a reference category. Because of this, odds ratios focus on the relative difference between a category of interest and a baseline category more than point estimates of response propensities provided by the survey managers. As a result, if the ordering and relative differences are accurate, that may be sufficient for generating relatively useful priors.

Associated with this is the fact that continuous variables were queried about on a linear scale, while the logistic regression modeling assumes a log-odds scale. For categorical variables this transformation is straightforward, since there is only

a fixed set of options for the categorical variable to take; for continuous covariates, however, extrapolations outside of the specific values considered lead to different predictions. Thus, if an expert suggests that an additional contact attempt increasing the probability of a successful contact from 5% from a 24% baseline, this yields a beta parameter of 0.26; thus five contact attempts increase the odds of contact to 54%, instead of the 49% on the linear scale, and to 81% after transformation from the log-odds scale for 10 contact attempts, vs. 74% on the original linear scale. Hossack, Hayes and Barry (2017) have proposed eliciting priors at a series of quantiles of the continuous predictor values in order to better approximate the log-odds transformation; we leave this as a future extension.

An iterative process to address these issues is difficult to carry out without collaboration with the targeted experts and may not be possible in all situations. However, if it is possible to first validate a questionnaire with some experts, keeping in mind the potential biases like overconfidence and anchoring biases, the resulting questionnaire may have more predictive power. Similarly, the SHELF method, proposed by O'Hagan (2019) relies on a significant amount of interaction with the experts throughout the elicitation process in order to elicit a probability distribution form each expert. While this method can be highly informative, providing both a point estimate and a measure of uncertainty for each expert's opinion, the number of items in our questionnaire would not have allowed for this level of individual interaction.

We also used the variability in the point estimates across our sample of experts to determine the variability in the prior distribution. This simplified the task of constructing the prior, since the experts were required only to supply point estimates, not estimates of uncertainty. This required a relatively large sample size of experts compared to many such elicitation studies. It also allowed us to take advantage of the Central Limit Theorem to utilize a normally-distributed prior, which in turn allowed more direct comparisons with West et al. (2019); alternatively, more heavy-tailed priors (e.g., t-distributions with small degrees of freedom) could be used. We did not rescale the prior to account for this sample size; one could construct a prior based on a "pseudo-sample size" of $m$ by multiplying $SE\left(\widehat{\widehat{\beta}}_{jk}\right)$ in (4) by $\sqrt{n/m}$ (that is, standard deviation of the arithmetic mean by the square root of $m$ rather than the square root of the actual number of respondents). Alternatively, one could elicit estimates of uncertainty as well as point estimates from the expert sample, and use information for both the direct elicitation and the sampling variability to construct the variance of the prior; we leave this to future research.

A limitation of our approach is that we used historical data to determine the key covariates to include in our survey of experts. We did this in order to make a fair comparison with historical data in our analysis, but in practice one might at best have data available from other studies with greater or lesser degrees of similarity. Indeed, one might have no historical data whatsoever from which to build a

propensity model, in which case one would have to rely on experts' opinion about potentially predictive items to develop an effective model for response propensity. As noted in Section 5, querying experts for the key covariates may have advantages over model selection, even if historical data is available from similar studies.

Finally, it is important to elicit expert opinion from appropriate individuals, based on the survey characteristics. Experts at ISR were identified through discussions with survey managers to identify appropriate individuals. At the Census Bureau, we worked with senior leadership in the Field Directorate to identify the two "most knowledgeable" survey managers in each of the six regional offices. This provided geographic coverage over the entire country and, we hoped, significant experience in demographic surveys that could be translated into priors for response propensity prediction. We did not include any other requirements in our identification of survey managers for interview. After collecting responses, we found that survey experience ranged anywhere from '0-4 years' to '15 or more years', and we found potential correlations between experience and predictions of attempt-level response rates predictions for some covariates. Due to the small sample size, we cannot conclude that these correlations are meaningful. However, it is useful to consider whether additional requirements would be useful when identifying experts. Relevant experience, either with respect to survey topic (e.g., health, education, etc.), operations (e.g., multimode vs. in-person interviewer-administered), or other characteristics, may lead to more informative expert opinion for incorporating into priors.

# References

Axinn, W.G., Link, C.F., & Groves, R.M. (2011). Responsive survey design, demographic data collection, and models of demographic behavior. *Demography, 48,* 1127-1149.

Boulet, S., Ursino, M., Thall, P., Landi, B., Lepère, C., Pernot, S., Burgun, A., Taieb, J., Zaanan, A., Zohar, S., & Jannot, A.-S. (2019). Integration of elicited expert information via a power prior in Bayesian variable selection: Application to colon cancer data. *Statistical Methods in Medical Research.* doi: 10.1177/0962280219841082.

Chapman, C. (2014). National Center for Education Statistics Adaptive Design Overview, Federal Committee on Statistical Methodology Conference, Washington, DC, December 16th.

Coffey, S., Reist, B., & Miller, P. V. (2019). Interventions On-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates, *Journal of Survey Statistics and Methodology.* doi: 10.1093/jssam/smz026

Couper, M.P. (2000). Usability Evaluation of Computer-Assisted Survey Instruments. *Social Science Computer Review, 18* (4), 384-396.

Couper, M. (2017). Birth and Diffusion of the Concept of Paradata (in Japanese – translated by W. Matsumoto). *Advances in Social Research, 18*, 14-26. Retrieved from: http://jasr.or.jp/english/JASR_Birth%20and%20Diffusion%20of%20the%20Concept%20of%20Paradata.pdf

Dallow, N., Best, N., & Montague, T.H. (2018). Better decision making in drug development through adoption of formal prior elicitation. *Pharmaceutical Statistics, 17*, 301-316. doi: 10.1002/pst.1854

Gelman, A., Carlin, J.B., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis*. Boca Raton: Chapman Hall.

Groves, R.M. & Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*, 439-457.

Hampson, L. V., Whitehead, J., Eleftheriou, D., & Brogan, P. (2014). Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine, 33*, 4186-4201. doi: 10.1002/sim.6225

Hosack, G. R., Hayes, K. R., & Barry, S. C. (2017). Prior elicitation for Bayesian generalised linear models with application to risk control option assessment. *Reliability Engineering & System Safety, 167*, 351-361.

Laflamme, F., & Karaganis, M. (2010). Development and implementation of responsive design for CATI surveys at Statistics Canada, European Quality Conference, Helsinki. Retrieved from: https://www.researchgate.net/publication/228583181_Implementation_of_Responsive_Collection_Design_for_CATI_Surveys_at_Statistics_Canada

Lepkowski, J.M., Mosher, W.D., Groves, R.M., West, B.T., Wagner, J., & Gu, H., (2013). Responsive design, weighting, and variance estimation in the 2006-2010 National Survey of Family Growth. *Vital and Health Statistics, 2*(158).

Mason, A. J., Gomes, M., Grieve, R., Ulug, P., Powell, J. T., & Carpenter, J. (2017). Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the IMPROVE trial. *Clinical Trials, 14*(4), 357-367. doi: 10.1177/1740774517711442

O'Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. *The American Statistician, 73*, 69-81.

Peytchev, A., Rosen, J., Riley, S., Murphy, J., & Lindblad, M. (2010). Reduction of Nonresponse Bias through Case Prioritization, *Survey Research Methods, 4*, 21-29.

Roberts, C., Vandenplas, C., & Stahli, M.E. (2014). Evaluating the impact of response enhancement methods on the risk of nonresponse bias and survey costs, *Survey Research Methods, 8*, 67-80.

Rosen, J.A., Murphy, J., Peytchev, A., Holder, T., Dever, J.A., Herget, D.R., & Pratt, D.J. (2014). Prioritizing low-propensity sample members in a survey: Implications for nonresponse bias. *Survey Practice, 7*(1).

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., & Wagner, J. (2018). A Bayesian Analysis of Design Parameters in Survey Data Collection. *Journal of Survey Statistics and Methodology, 6*, 431-464.

Schouten, B., Peytchev, A., & Wagner, J. (2017). *Adaptive Survey Design*. Boca Raton, Florida: CRC Press.

Schouten, B., Calinescu, M., & Luiten, A. (2013). Optimizing quality of response through adaptive survey design, *Survey Methodology, 39*, 29-58.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester ; Hoboken, NJ, John Wiley & Sons.

Thompson, J., & Kaputa, S. (2017). Investigating adaptive non-response follow-up strategies for small businesses through embedded experiments. *Journal of Official Statistics, 33*(3), 835-856.

Wagner, J. & Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology, 2*, 323-342.

Wagner, J., West, B. T., Guyer, H., Burton, P., Kelley, J., Couper, M. P., & Mosher, W. D. (2017). The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth. In P. P. Biemer, E. de Leeuw, S. Eckman et al., *Total Survey Error in Practice*. New York, Wiley.

Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G., & Kruger-Ndiaye, S. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics, 28*, 477-499.

West, B.T., & Groves, R.M. (2013). The PAIP Score: A propensity-adjusted interviewer performance indicator. *Public Opinion Quarterly, 77*, 352-374.

West, B.T. (2013). An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society, Series A, 176*, 211-225.

West, B.T., Wagner, J., Gu, H., & Hubbard, F. (2015). The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology, 3*, 240-264.

West, B.T., Wagner, J., Coffey, S., & Elliott, M.R. (2019). The Elicitation of Prior Distributions for Bayesian Responsive Survey Design: Historical Data Analysis vs. Literature Review. Retrieved from https://arxiv.org/ftp/arxiv/papers/1907/1907.06560.pdf.

# Appendix

*Table A1* Significant predictors of screener response propensity in the final discrete time logit model for call-level data from the eight most recent quarters, after applying backward selection (n = 119,981 calls; Nagelkerke pseudo R-squared = 0.09; AUC = 0.66).

| Predictor | Coefficient | Standard Error |
|---|---|---|
| Intercept | -2.56 | 0.32 |
| Mail Delivery Point Type: Missing | 0.08 | 0.03 |
| Mail Delivery Point Type: A | 0.03 | 0.02 |
| Mail Delivery Point Type: B | -0.04 | 0.03 |
| Mail Delivery Point Type: C | -0.09 | 0.03 |
| Interviewer-Judged Eligibility: Missing | 2.46 | 0.10 |
| Interviewer-Judged Eligibility: No | 0.63 | 0.07 |
| Segment Listed: Car Alone | 0.03 | 0.02 |
| PSU Type: Non Self-Representing | 0.06 | 0.03 |
| PSU Type: Self-Representing (Not Largest 3 MSAs) | 0.03 | 0.03 |
| Previous Call: Contact | 3.97 | 0.28 |
| Previous Call: Different Window | -0.12 | 0.02 |
| Previous Call: Building Ever Locked | 0.32 | 0.05 |
| Previous Call: Building Locked | 2.16 | 0.14 |
| Previous Call: Strong Concerns Expressed | 0.26 | 0.04 |
| Previous Call: No Contact | 2.26 | 0.13 |
| Previous Call: Other Contact, No Concerns Expressed | -1.35 | 0.25 |
| Previous Call: Concerns Expressed | -1.58 | 0.26 |
| Previous Call: Soft Appointment | -1.03 | 0.30 |
| Previous Call: Call Window Sun.-Thurs. 6pm-10pm | 0.07 | 0.03 |
| Previous Call: Call Window Fri.-Sat. 6pm-10pm | 0.08 | 0.02 |
| No Access Problems in Segment | -0.05 | 0.02 |
| Evidence of Other Languages (not Spanish) | -0.09 | 0.03 |
| Census Division: G | -0.14 | 0.03 |
| Census Division: B | -0.32 | 0.03 |
| Census Division: D | -0.22 | 0.03 |
| Census Division: H | -0.24 | 0.03 |
| Census Division: C | -0.20 | 0.03 |
| Census Division: F | -0.27 | 0.04 |
| Census Division: E | -0.20 | 0.03 |
| Census Division: A | -0.19 | 0.04 |
| Contacts: None | -0.68 | 0.24 |
| Contacts: 1 | -0.54 | 0.22 |

| Predictor | Coefficient | Standard Error |
|---|---|---|
| Contacts: 2 to 4 | -0.42 | 0.19 |
| Segment Domain: <10% Black, <10% Hispanic | -0.04 | 0.02 |
| Segment Domain: >10% Black, <10% Hispanic | -0.04 | 0.02 |
| Segment Domain: <10% Black, >10% Hispanic | 0.01 | 0.03 |
| Percentage of Segment Non-Eligible (Census Data) | -0.01 | <0.01 |
| Interviewer-Estimated Segment Eligibility Rate | -0.55 | 0.12 |
| Interviewer-Estimated Household Eligible | -0.09 | 0.02 |
| Segment Type: All Residential | 0.04 | 0.02 |
| Log(Number of Calls Made) | -0.60 | 0.03 |
| Log(Number of Calls Made) x No. Prev. Contacts | -0.04 | 0.01 |
| CML* HoH Age: 35-64 | -0.12 | 0.02 |
| CML Adult Count: Missing | -0.13 | 0.04 |
| CML Adult Count: 1 | -0.09 | 0.03 |
| CML Adult Count: 2 | 0.01 | 0.03 |
| CML Asian in HH: Missing | 0.21 | 0.04 |
| CML Asian in HH: No | 0.20 | 0.05 |
| CML HoH Gender: Missing | -0.03 | 0.02 |
| CML HoH Gender: Female | -0.01 | 0.02 |
| CML HoH Income: $35k-$70k | 0.12 | 0.02 |
| CML HoH Income: less than $35k | 0.14 | 0.02 |
| CML HH Own/Rent: Missing | -0.06 | 0.03 |
| CML HH Own/Rent: Owned | -0.02 | 0.02 |
| CML Age of 2nd Person: Missing | -0.13 | 0.03 |
| CML Age of 2nd Person: 18-44 | -0.15 | 0.03 |
| No Respondent Comments | 0.08 | 0.04 |
| Non-Contacts: None | -0.51 | 0.08 |
| Non-Contacts: 1 | -0.25 | 0.05 |
| Non-Contacts: 2-4 | -0.03 | 0.03 |
| Occupancy Rate of PSU | -0.26 | 0.10 |
| Respondent Other Concerns | 0.18 | 0.06 |
| Physical Impediment to Housing Unit: Locked | -0.35 | 0.03 |
| Day of Quarter | 0.01 | <0.01 |
| Respondent Concerns Expressed: None | -1.25 | 0.15 |
| Respondent Concerns Expressed: Once | 0.15 | 0.09 |
| Single Family Home / Townhome | -0.22 | 0.03 |
| Structure with 2-9 Units | -0.29 | 0.04 |
| Structure with 10+ Units | -0.21 | 0.04 |
| Respondent Concern: Survey Voluntary? | -0.46 | 0.15 |
| Respondent Concern: Too Old | 0.60 | 0.15 |

* CML denotes that the variable came from a commercial data source.

*Table A2* Normal Prior Definitions, $\left(\hat{\bar{\beta}}_{jk}, SE\left(\hat{\bar{\beta}}_{jk}\right)\right)$, for all predictors included in the NSFG response propensity model described in Section 3.2. The table notes which categories served as reference categories in the prior generation process, and also notes how many responses (out of a maximum of 20) that we received for each category.

| Questions and Categories | All Respondents (max n = 20) | | |
|---|---|---|---|
| | Count of Responses | Mean Beta | StdErr Beta |
| *Gender of Primary Householder (vs. Male)* | | | |
| Female | 20 | 0.336 | 0.063 |
| Missing | 14 | -0.465 | 0.257 |
| *Age of Primary Householder (vs. 50 or Over)* | | | |
| < 50 | 20 | -0.370 | 0.108 |
| Missing | 15 | -0.831 | 0.293 |
| *Number of Adults in HH (vs. 2 or More)* | | | |
| 1 | 20 | 0.066 | 0.198 |
| Missing | 12 | -0.732 | 0.219 |
| *Race/Ethnicity of Primary Householder (vs. Asian)* | | | |
| White | 18 | 0.532 | 0.121 |
| Black | 18 | -0.031 | 0.173 |
| Hispanic | 18 | -0.118 | 0.112 |
| Other | 13 | -0.348 | 0.233 |
| Missing | 12 | -0.326 | 0.292 |
| Household Income Effect | | | |
| +$10,000 | 17 | 0.466 | 0.235 |
| *Masked Census Division (vs. Region I)* | | | |
| G | 14 | 0.020 | 0.129 |
| B | 14 | -0.205 | 0.138 |
| D | 14 | 0.041 | 0.141 |
| H | 14 | 0.060 | 0.161 |
| C | 14 | 0.133 | 0.170 |
| F | 15 | 0.294 | 0.150 |
| E | 15 | 0.057 | 0.145 |
| A | 16 | -0.050 | 0.192 |
| *Race/Ethnicity Sampling Domain (vs. > 10% Black, > 10% Hispanic)* | | | |
| < 10% Black, < 10% Hispanic | 16 | 0.696 | 0.202 |
| > 10% Black, < 10% Hispanic | 16 | 0.535 | 0.132 |
| < 10% Black, > 10% Hispanic | 16 | 0.364 | 0.143 |
| *Access Problems (vs. Other)* | | | |
| Locked Buildings/Gated Communities | 19 | -0.687 | 0.190 |

| Questions and Categories | All Respondents (max n = 20) | | |
|---|---|---|---|
| | Count of Responses | Mean Beta | StdErr Beta |
| Seasonal Hazardous Conditions | 18 | -0.418 | 0.153 |
| Unimproved Roads | 17 | 0.267 | 0.164 |
| None | 10 | 1.091 | 0.189 |
| *Evidence of Non-English Languages (vs. No)* | | | |
| Yes | 15 | -0.725 | 0.163 |
| *Neighborhood Age Effect* | | | |
| 10 years older than national average | 17 | 0.520 | 0.099 |
| *Occupancy Rate Effect* | | | |
| 10% increase in occupancy rates | 16 | 0.187 | 0.170 |
| *PSU Type (vs. Major Metropolitan Area)* | | | |
| Minor Metropolitan Area | 18 | 0.155 | 0.155 |
| Not Metropolitan | 17 | 0.398 | 0.158 |
| *Listing Procedure (vs. On Foot Alone)* | | | |
| On Foot With Someone | 11 | 0.787 | 0.607 |
| In a Car Alone | 11 | -0.066 | 0.135 |
| In a Car With Someone | 11 | 0.795 | 0.614 |
| *Structure Type (vs. Other)* | | | |
| Single Family Home | 5 | 1.172 | 0.567 |
| Structure with 2-9 Units | 5 | 0.788 | 0.602 |
| Structure with 10+ Units | 5 | 0.600 | 0.617 |
| Mobile Home | 5 | 0.728 | 0.462 |
| *Delivery Type (vs. Other)* | | | |
| Curbline | 3 | 0.917 | 0.590 |
| Neighborhood Delivery Collection Box | 3 | 0.199 | 0.289 |
| Central | 3 | 0.069 | 0.384 |
| Missing | 3 | 0.000 | 0.000 |
| *Physical Impediments (vs. Other)* | | | |
| Locked Entrance | 19 | -0.096 | 0.206 |
| Doorperson or Gatekeeper | 19 | -0.627 | 0.117 |
| Access controlled via Intercom | 19 | -0.371 | 0.106 |
| None | 14 | 1.076 | 0.155 |
| *Attempt-Level Concerns Expressed (vs. No Concerns)* | | | |
| Concerns Expressed on Previous Attempt | 17 | -1.347 | 0.434 |
| Concerns Expressed Not on Previous but Prior Attempt | 17 | -1.451 | 0.244 |
| Strong Concerns Ever Expressed | 15 | -2.228 | 0.593 |
| *Attempt-Level Contact (vs. Never Contacted)* | | | |
| Contacted at Previous Attempt | 15 | 1.367 | 0.329 |
| Not Previous but Prior Contact | 15 | 1.009 | 0.298 |

| Questions and Categories | All Respondents (max n = 20) | | |
|---|---|---|---|
| | Count of Responses | Mean Beta | StdErr Beta |
| *Contact Observations (vs. Other)* | | | |
| Ever Said „Too Old“ | 14 | -0.532 | 0.336 |
| Comment re: Voluntary Nature of Survey | 17 | 0.335 | 0.489 |
| Any Other Comments | 14 | 0.118 | 0.182 |
| Never Made Comment | 13 | 0.325 | 0.205 |
| *Day of Field Period Effect* | | | |
| Change in RR for Each Day of Field Period | 12 | 0.213 | 0.078 |
| *Call Window (vs. Weekday Day)* | | | |
| Weekday Evening | 19 | 1.203 | 0.193 |
| Weekend Day | 19 | 1.052 | 0.166 |
| Weekend Evening | 19 | 0.426 | 0.220 |
| *Ever Requested Call-Back/Soft Appointment (vs. No)* | | | |
| Yes | 18 | 0.564 | 0.339 |
| *Concatct Attempt Effect* | | | |
| Change in RR for Each Additional Contact | 17 | -0.058 | 0.109 |
| *Contact*Contact Interaction Effect* | | | |
| Change in RR for Each Add‘l Call*Contact | 13 | 0.177 | 0.228 |

methods, data, analyses | Vol. 14(2), 2020, pp. 195-214 DOI: 10.12758/mda.2020.06

# Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order

*Carsten Sauer*[1], *Katrin Auspurg*[2] *& Thomas Hinz*[3]

[1] *Department of Political and Social Sciences, Zeppelin University Friedrichshafen*

[2] *Department of Sociology, LMU Munich*

[3] *Department of History and Sociology, University of Konstanz*

**Abstract**

Multi-factorial survey experiments have become a well-established tool in social sciences as they combine experimental designs with advantages of heterogeneous respondent samples. This paper investigates three under-researched design features: how to present vignettes (running text vs. table), how to measure responses (rating vs. open scale), and how to sort vignettes (random vs. extreme-cases-first, to prevent censored responses). Experiments were conducted in a 2 x 2 x 2 between-subject design with 408 university students rating decks à 20 vignettes. Analyses of 7,895 ratings showed no differences of whether vignettes were presented as running texts or tables. Open scales revealed more measurement problems, e.g., missing values, than rating scales. Finally, vignettes presented randomly sorted produced similar results compared to sorting extreme vignette cases first. Recommendations based on the findings are to use random orders of vignettes and rating scales. Table vignettes provide an alternative to text vignettes but should be further evaluated with heterogeneous samples.

*Keywords*: Multi-factorial survey, vignette presentation, response scale, vignette order, ceiling effects

Multi-factorial survey experiments have become a well-established tool in the social sciences, mostly because they combine experimental design features (i.e. randomization) with the advantages of heterogeneous respondent samples (i.e. large and/or random samples that enable the estimation of heterogeneous treatment effects). In these survey experiments participants respond to descriptions of hypothetical objects or situations (vignettes). Within the vignettes, factors (dimensions) vary experimentally in their levels. The experimental variation allows an analysis of dimensions' causal influence on the responses (normative judgments or hypothetical decisions). At the same time, as the experiment is embedded in a survey, it is a tool to reach heterogeneous respondent samples and to analyze differences in attitudes or behavioral intentions across social groups. During the last years increasing numbers of studies have been published indicating that multi-factorial survey experiments became more and more a standard tool in social sciences (Auspurg & Hinz, 2015; for multifactorial, "conjoint" survey experiments in political sciences: Hainmueller, Hangartner, & Yamamoto, 2015).

Whenever implementing such experiments, researchers make decisions about multiple design features. Previous research focused on the complexity (number of dimensions and vignettes; see Auspurg, Hinz, & Liebig, 2009; Sauer, Auspurg, Hinz, & Liebig, 2011), sampling techniques (Atzmüller & Steiner, 2010; Dülmer, 2007, 2016), survey mode (Weinberg, Freese, & McElhattan, 2014), methods of data analyses (Hox, Kreft, & Hermkens, 1991), and external validity (Hainmueller et al., 2015; Petzold & Wolbring, 2019). Our study extends this literature by inves-

*Acknowledgements*

The data collection reported in this paper was supported by a grant funded by the German Research Foundation (DFG) (Hi 680/4-1). The paper was presented at the workshop "A perfect match? Comparative Political Economy and conjoint analysis" at the University of Zurich in 2017. We thank the organizers, in particular Silja Häusermann, and the participants, especially Devin Caughey, for their valuable comments and suggestions. Moreover, we thank participants of the colloquium of the Department of Sociology in October 2016 at Radboud University for many helpful comments. For contributing to this paper, Thomas Hinz was supported by EXC2035 "The Politics of Inequality". Carsten Sauer acknowledges support by the Dutch Research Council (NWO, Veni grant number: 4510-17-024).

*Data Note*

This article uses data from the project "The Factorial Survey as a Method for Measuring Attitudes in General Population Surveys". Replication files (Stata do-files and the used data) can be found on the following webpage: https://dx.doi.org/10.7802/2011

*Direct correspondence to*

Carsten Sauer, Lehrstuhl für Soziologie mit Schwerpunkt Sozialstrukturanalyse, Fakultät für Staats- und Gesellschaftswissenschaften, Zeppelin Universität, Am Seemooser Horn 20, 88045 Friedrichshafen
E-mail: carsten.sauer@zu.de

tigating the effects of three fundamental design features on the data quality which received little attention so far: first, presenting the vignettes in a running text vs. table format; second, using open response scales vs. rating scales with closed ends; and third, a random or systematic (extreme-cases-first) order of the vignettes presented to the respondents. The first two design features are crucial for all researchers in the field as they must decide how to present information and choose (at least) one answering scale. The third question about the vignette order is additionally important for the large bulk of applications with multiple vignettes per respondent: Researchers typically ask respondents to evaluate several (e.g., 10 or 20) vignettes (for a review of applications, see Wallander, 2009). As we will explain in more detail below, in these cases ordering the vignettes in systematic (instead of random) way is seen as a promising tool to avoid censored responses, but there are so far no empirical evaluations.

In the literature, there are some guidelines for the construction of multifactorial experiments to gather most reliable and valid results (Auspurg & Hinz, 2015; Jasso, 2006; Sauer, Auspurg, Hinz, Liebig, & Schupp, 2014). The findings of our study provide additional insights as so far only few studies contrasted a text and tabular format (Shamon, Dülmer, & Giza, 2019), an open and a rating scale (Auspurg & Hinz, 2015), and/or different vignette orders within the same experimental design.

## Background: Why Should the Design Features Make a Difference?

*Presentation Style.* Vignettes used in multi-factorial survey experiments typically describe hypothetical situations or persons by a running text, i.e. a paragraph of one or several full sentences (see, Auspurg & Hinz, 2015, pp. 69-72). By doing so, the vignettes describe short scenarios close to 'real-life-stories,' which is seen as a main advantage of this presentation style. Moreover, it allows for a very subtle, indirect question format that can be useful to investigate sensitive topics (Auspurg, Hinz, Liebig, & Sauer, 2015). An alternative style would be a table format that only shows the dimensions and levels and avoids additional text. This presentation style is frequently used in conjoint studies and choice experiments, i.e. multi-factorial survey experiments that prevail in marketing research and economics. Critical about this tabular presentation style might be the more abstract question format which is not embedded in a story. Further possible limitations exist with respondents more likely using heuristics or being more prone to social desirability bias when the dimensions are presented more evidently in tables instead of being 'hidden' in smooth stories. However, there are also lots of advantages of tables: The format might minimize respondents' cognitive burden by reducing the reading task. Information presented

in tables can be assessed faster and should therefore economize on survey time. Additionally, table formats provide an appealing alternative to running text if one wants to randomize the order of the dimensions to neutralize potential effects of the dimension order (such as primacy and recency effects, see Auspurg & Jäckle, 2017). Vignette dimensions can more easily be rotated in a tabular format, as the order is no longer specific to the syntax of a language. In text vignettes, moreover, respondents might simply overlook some dimensions, which would obviously invalidate results gained by such experiments. Thus, even though running texts are mostly used in multi-factorial survey experiments so far, table formats may be a versatile alternative. So far, one study investigated differences between tabular vignettes and text vignettes using an online quota sample (Shamon et al., 2019) and finds no differences between the two methods regarding response inconsistency and processing time but more missing values (including refusals to answer any vignette at all) for text vignettes especially for respondents with lower educational degrees.

*Response Scales.* There are several ways to measure the responses to the vignette stimuli (see Auspurg & Hinz, 2015, pp. 64-67; Wallander, 2009). We tested the most frequently used response scales of vignette studies in the social sciences, an ordered rating scale (in our case an 11-point scale) against an open scale, also known as magnitude scale (Jasso, 2006; Sauer et al., 2011). The advantage of rating scales is that they are easily accessible for respondents as they are frequently used in various types of survey questions and, therefore, represent a standard tool of survey research. However, obviously, the range of values is restricted by the predefined minimum and maximum of such a scale. For this reason, ceiling effects might occur: In particular, when respondents have to rate multiple vignettes, they might not be able to express a more nuanced judgement that is located between to scale points or that goes beyond the scale's minimum or maximum. The resulting censored responses would lead to a systematic underestimation of the effects of vignette dimensions (i.e. there is a lower statistical power to detect the vignette dimensions' impact). Open (magnitude) scales that have no limits are deemed to overcome such ceiling effects and also to provide more fine-grained, metric values (Jasso, 2006). The drawback is that these scales likely cause a higher cognitive burden for the respondents. Open (magnitude) scales have been frequently used and recommended for multi-factorial survey experiments and conjoint analyses (for an overview, see Liebig, Sauer, & Friedhoff, 2015), but tests of their reliability are missing. (To best of our knowledge, the only systematic evaluation for multi-factorial survey experiments exists with a small marketing survey, a conjoint analysis, with 100 respondents in the U.S.; see Teas 1987.)

*Vignette Order.* The use of a random order of vignettes allows neutralizing possible effects of a fixed vignette order (such as carry-over, learning or fatigue effects). However, to avoid ceiling effects, some authors alternatively presented the vignettes in a systematic order, starting with the most extreme vignette cases. The

reason for this recommendation is that beginning with the vignettes likely to provoke the most extreme reactions could help to calibrate respondents regarding the end points of closed-ended rating scales (Auspurg & Hinz, 2015). Yet one drawback is that the researchers must decide which vignettes respondents may perceive as extreme cases. Systematic comparisons of both orders are lacking.

*Interactions between the design features.* Although it is not the core question of this study, our orthogonal, multi-factorial experimental design also allows us to test interaction effects between all three design features. The vignette order and response scales might have a different impact for tabular vignettes with a clear-structured presentation format compared to text vignettes, where respondents might be less aware of all dimensions. Similarly, the use of an extreme-case-first order might be especially effective in combination with closed-ended rating scales that are more prone to ceiling effects.

## Data and Methods

We fully crossed all three design features (text/tables, response scales, and vignette order), leading to a 2 x 2 x 2 between-subject experiment (the between-subject design was chosen to not distract the respondents with changing scales or presentation styles). The substantive issue of the factorial survey module was the fairness of earnings of hypothetical full-time employees. The analysis sample consisted of 408 bachelor students of social sciences, 177 men and 231 women. All participants were recruited in 2008 in social science courses at 27 German universities and then randomly allocated to one of the 8 different experimental cells.[1] Depending on the local conditions, respondents could answer to the online survey (CASI) either during their course or afterwards in their free time. The questionnaire started with some socio-demographic questions, e.g., about the field of studies. The vignette module started with an introductory screen that provided shortly some general information on the hypothetical employees that was held constant for all vignette persons, such as their weekly working hours (40 hours). The following vignette module included 20 vignettes for each respondent. Table 1 provides the realized numbers of observations (rated vignettes) and number of participants per experimental cell.

In the vignettes, information on hypothetical employees participating in the German labor market was presented. The 8 dimensions (including the gross earnings) were selected close to prior factorial survey studies in the substantive field

1 The data collection was part of a larger project that investigated multiple methodological issues of multi-factorial survey experiments such as effects of the number of dimensions and levels, mode effects, and the reliability of measurement. Participating universities were recruited via personal contacts to the PIs.

*Table 1* Number of Vignettes and Respondents (in Parentheses) per Experimental Cell

| | Type of scale | | | | |
|---|---|---|---|---|---|
| | Rating scale | | Open scale | | |
| Presentation of dimensions | Random order | Extreme cases first | Random order | Extreme cases first | Total |
| Text | 1,087 (56) | 1,044 (53) | 916 (47) | 839 (45) | 3,886 (201) |
| Table | 1,159 (58) | 1,099 (55) | 886 (47) | 865 (47) | 4,009 (207) |
| Total | 2,246 (114) | 2,143 (108) | 1,802 (94) | 1,704 (92) | 7,895 (408) |

*Table 2* Vignette Dimensions and their Levels

| # | Dimensions | (Number of) levels |
|---|---|---|
| 1 | Age | (4) 30, 40, 50, 60 years |
| 2 | Sex | (2) male, female |
| 3 | Degree | (3) without degree, vocational degree, university degree |
| 4 | Occupation | (10) unskilled worker, door(wo)man, engine driver, clerk, hairdresser, social worker, software engineer, electrical engineer, business manager, medical doctor |
| 5 | Experience | (2) short on, much |
| 6 | Tenure | (2) entered recently, entered a long time ago |
| 7 | Children | (5) no child, 1 child, 2, 3, 4 children |
| 8 | Earnings | (10) values from 500 to 15.000 Euros |

(e.g., Jasso & Rossi, 1977; Shepelak & Alwin, 1986). Table 2 shows all dimensions and levels. Each vignette was presented on a single screen page. The task for the respondents was to assess the justice of the gross earnings. Respondents had the possibility to skip evaluations (no forced evaluations) and to return to vignettes evaluated before if they wanted to change their evaluation. About 92 percent of vignettes were visited only once, thus, respondents did not change their ratings. In 8 percent of the cases people went back to previous screens to change their judgements. Screenshots of some exemplary vignettes are provided in Online-Appendix, Part A.

We used a sample of vignettes as the full-factorial of all combinations of dimension levels would yield 48,000 vignettes. Our selection of 240 vignettes (12 decks à 20 vignettes) was based on the D-efficiency criterion (Kuhfeld, Tobias,

& Garratt, 1994). With this sampling method, it is possible to find a selection of vignettes in which correlations between dimensions are minimized (overall and within the different decks; criterion of orthogonality). At the same time, it is ensured that all levels of each dimension appear similarly often (criterion of level balance). Both criteria ensure that one receives a sample that allows to estimate coefficients efficiently and unbiased. Illogical and very implausible combinations were excluded, like medical doctors without a university degree.[2] (For a detailed description of the sampling method and comparisons with alternative designs, see Auspurg & Hinz, 2015).

The experimental manipulations were set-up as follows: The running text vignettes were programmed as shown in the sample vignette presented in Figure A1 in the Appendix A. The table format was programmed with 4 rows and 2 columns showing the dimensions and their levels (Figure A2_1 and A2_2). In these table vignettes, the order of dimensions was fixed to have equivalent conditions as in the text vignettes.

The answering scales were programmed in two versions with an 11-point rating scale versus an open (magnitude) scale. The rating scale had the standard format used in previous vignette studies with the scale running from -5 (unfairly too low) over zero (fair) up to +5 (unfairly too high). For the magnitude scale, we implemented a design very similar to that described in a prominent instruction on factorial surveys (Jasso, 2006).[3] This answering scale followed a three-step procedure (shown in Figure A2_1 and A2_2) as it is recommended in the literature (Jasso, 2006). First, respondents evaluated if the earnings of the vignette person were just or unjust. If respondents rated the earnings to be just, they approached to the next vignette. If respondents evaluated the earnings to be unjust, they answered in a second step whether the earnings were too high or too low. In a third step the participants were asked to specify the amount of injustice. Respondents could use their own unrestricted continuum of numbers that express their perception of injustice best for this evaluation step. Based on the insights of psychophysics (Stevens, 1975) these numbers are deemed to be metric evaluations. To have a reference point for these evaluations across respondents, a calibration vignette, which was the same for all respondents, was added in front of the vignette decks in the magnitude-split; i.e. all respondents first had to evaluate this calibration vignette (see Jasso (2006) for an in-depth description of this approach). For data analyses, these three response variables were transformed into one joint measurement following Jasso (2006): First, the ratings were combined within one numeric scale with zeros describing perfect justice, negative numbers describing under-reward and positive numbers describing

2 Plausible interaction terms have been orthogonalized (Resolution-IV-design). The D-efficiency of the 240 vignettes sampled was 91.

3 The method is based on psychophysics (Stevens, 1975) and has been applied in many factorial survey studies (for an overview in the justice literature, see Liebig et al., 2015).

over-reward. Second, the number continuums used by different respondents were calibrated by dividing these numbers by the rating of the calibration vignette.[4]

Regarding the variation of the vignette order, respondents evaluated in the first condition vignettes that were ordered randomly. For each respondent, the random order of the 20 vignettes in their deck was generated by a random number generator (we used the statistical software Stata). The second condition was an extreme-cases-first order. In this split, first, again for each respondent a random order of the twenty vignettes was generated. After the randomization, the order was manipulated by moving the two most extreme vignette cases (high underpayment and high overpayment) to the beginning of the vignette module. The driving dimensions for the selection of these extreme cases were the "gross earnings" and "occupation": We selected the two vignette cases that showed the highest (lowest) earnings given what is common in Germany for the respective occupations. To determine these cases' earnings, we used official information about the actual earnings by occupation from labor market data in Germany.[5] Information on earnings per occupation was chosen because existing surveys (and also our survey) showed that respondents in Germany account in their justice evaluations very strongly for what people realistically earn in different occupations. Therefore, these two vignettes could be expected to evoke extreme ratings in both directions (over- and underpaid). Putting them first is thought to lessen ceiling effects in later judgments of less extreme vignettes (Garret, 1982; O'Toole, Webster, O'Toole, & Lucal, 1999).[6]

*Data Analyses*. Data were analyzed using linear multi-level (random-intercept) regressions, with vignette evaluations at level 1 and respondents at level 2. The outcome variable was the vignette ratings of the respondents. To make estimates based on the open (magnitude) scale comparable to those based on the rating scale, all ratings were z-standardized. As input variables we used the vignette dimensions described in Table 2. The dimensions "degree" and "occupation" were included as dummy sets.

To identify if design features affected the importance of different dimensions for the judgements, we chose the following strategy: For each experimental split, the 17 coefficients were interacted with a binary-indicator for the two design variants (text vs. table format, rating vs. open scale, random order vs. extreme cases

4 The calibration has the drawback that one needs valid values in these first judgments. In our study 11 respondents produced missing values and 9 respondents evaluated the first vignette as just (0) and could therefore not be used for the calibration.

5 When there were several extreme vignette earnings in a deck (i.e. vignette earnings were at least for two vignettes twice or even three times the mean actual earnings for this occupation) we additionally used information on the educational degrees to determine the two most "extreme" under-/overpaid vignette cases.

6 Extremely under-rewarded vignette persons were, e.g., medical doctors with meagre earnings; extremely over-rewarded vignettes persons were, e.g., unskilled workers with top-earnings.

first) to test for significant differences. Control variables included the respective other design features as well as respondent's sex and the university where the survey took place (26 dummies). We estimated linear multi-level regressions,[7] post-estimation tests were used to assess differences by our three experimental conditions. We employed $\chi^2$-tests for the null hypotheses that the interaction terms of vignette dimensions with the binary design indicator are (jointly) zero (this "omnibus" hypotheses test of that there are no differences at all is known as "Chow test", see Wooldridge, 2003). We report Sidak-adjusted *p*-values to account for multiple comparisons.

To check how design features affected response quality, we evaluated standard parameters to assess the response quality, such as the proportions of missing values with logistic regressions. In these analyses, we also explored two-way interactions between the different design features (e.g. between style of presentation and response scales). Moreover, we investigated response times and response consistency. General criteria to evaluate design features refer to the cognitive burden they impose on respondents. Obviously, the time respondents need to provide vignette evaluations serve as a proxy for the cognitive effort needed. We compare response times (measured during data collection for each of the 20 vignettes) by design splits and expect the scales to make a difference. For the analysis of response times we used median regression (Parente & Santos Silva, 2016). The consistency of responses is measured by another proxy, namely, the squared residuals following the procedure of Shamon et al. (2019) and Sauer et al. (2011). Lower values of squared residuals (given the same set of vignette dimensions for all respondents) are equal to a higher consistency in evaluations. While there are inter-individual differences (which are not at focus of this paper but see Auspurg, Hinz, & Liebig 2009) we assume again that the open scale is accompanied by less consistency. All data analyses were done with the statistical software Stata version 14.2 (StataCorp., 2013). The graphs were created with the user-written Stata ado *coefplot* (Jann, 2014).

# Results

Before we report the results of the methods experiments, we take a quick look at the substantive results to check their plausibility based on the empirical justice literature. Respondents' evaluations led to plausible effects of vignette dimensions on justice evaluations and were in line with prior factorial survey experiments in the field of pay fairness: E.g., vignette persons were considered as being the more

7 Note, we used a Generalized Least Square (GLS) estimation that leads to approximately similar results as Maximum Likelihood (ML) estimation but makes no assumption about the distribution of the unit-specific error term. The results reported here are not affected by the estimation algorithm (GLS or ML) and lead to the same results.

likely underpaid, the higher their educational degree, labor market experience, and occupational prestige; and the lower their gross earnings. The substantive findings of these regression results are presented in Appendix B.

## Effects on the Impact of Vignette Dimensions

What is more interesting for the study at hand: Did the results (effect sizes of dimensions) depend on the experimentally varied method features like the way vignettes were presented or had to be evaluated by respondents? Table 3 shows the differences across our three experimental splits (the underlying, substantive regression models and their interpretation are provided in Appendix B). Model 1 reports the results for table vs. text vignettes. The non-significant $\chi^2$-values indicate that there are no differences in the effects of vignette dimensions on respondents' judgements between the two presentation styles. Moreover, the insignificant joint test reported in the last row of the table suggests that the two design variants (text or tables) produce similar results. Model 2 shows the differences in coefficients for open vs. rating scales. Of the 8 dimensions, 5 were found to show significant differences between the two scales and the highly significant joint test at the bottom of the table also indicated that the two scales produced strikingly different results. This difference will be analyzed in more detail in the subsequent paragraph. Note, even with an alternative categorical coding of the dependent variables (with three categories: under-rewarded, fair, over-rewarded) differences remained (see Appendix C), meaning that differences were not driven by outliers of the open (continuous) scale. Model 3 focuses on the splits in which the order of the vignettes was varied. Results show that differences (interaction effects) – both being tested separately or jointly – are statistically insignificant. That is, we did not observe any significant differences between coefficients estimated with a random order of vignettes or with extreme cases first. This result remains stable also in case of restricting the analysis sample only to respondents who did not change previous ratings (92 percent of the sample).

## Response Quality of Response Scales and Vignette Orders

The analyses so far showed that only the choice of the answering scale had a significant impact on the regression results. The question follows, which scale performed better? Additional analyses revealed that the number of missing values was remarkably higher in evaluations made with the open scale than with the rating scale. Within the rating split, 4,389 vignettes were evaluated and 131 (2.9 %) vignettes were not. Within the magnitude split, 3,816 vignettes were rated and 344 (8.3 %)

*Table 3* Tests for Design Effects on the Impact of Vignette Dimensions

| | | M1 Presentation: table vs. text | M2 Open vs. rating scale | M3 Extreme cases vs. random |
|---|---|---|---|---|
| | df | $\chi^2$ | $\chi^2$ | $\chi^2$ |
| Experimental variation x sex | 1 | 3.170 | 0.399 | 0.004 |
| Experimental variation x age | 1 | 2.381 | 0.521 | 1.221 |
| Experimental variation x degree | 2 | 0.823 | 6.111* | 1.402 |
| Experimental variation x children | 1 | 0.219 | 5.716* | 2.319 |
| Experimental variation x experience | 1 | 0.386 | 10.454** | 0.095 |
| Experimental variation x tenure | 1 | 1.370 | 0.001 | 1.312 |
| Experimental variation x earnings | 1 | 0.003 | 75.107*** | 1.613 |
| Experimental variation x occupation | 9 | 9.356 | 37.828*** | 5.011 |
| Overall | 17 | 22.497 | 177.836*** | 17.766 |

*Notes*. Tests after multi-level (random-intercept) regressions with interaction terms; df: degrees of freedom of the respective vignette dimension; reference category M1: text vignettes; M2: rating scale; M3 random order; Controlled for further experimental manipulations, respectively, and respondents' sex and place of survey (26 dummies for the universities). N_vignettes = 7895; N_respondents = 408; Sidak-adjusted $p$-values; * $p < .05$; ** $p < .01$; *** $p < .001$.

vignettes were not.[8] This difference indicates that the respondents had more problems (or were less cooperative) with the open scale with its three-step rating procedure. Table 4 shows the coefficients of a logistic regression on the probability of missing values and reveals that missing values were only significantly more likely with open scales (Model 1). As shown in Models 2-4, there were also no significant interactions between the type of scale and presentation style or vignette order, indicating the open scale to be the main driver of missing values.

Besides the probability of missing values, the share of explained variance (overall $R^2$ in Stata) of the linear multiple regression model (see Appendix B2) – as another measure of response quality – was remarkably lower with the open scale ($R^2 = .11$) than with the rating scale ($R^2 = .51$) indicating that a lot of noise in the data collected with the open scale affected the precision of estimation.

8 Note: 8,680 potential judgments = 4,389 valid rating scale judgments + 131 missing rating scale judgments + 3,506 valid open scale judgements + 344 missing open scale judgments + 310 missings because of failed calibration. The analysis of missing values only includes missings (131 + 344) that were produced by the respondents. The actual missings for the analysis of the open scale split were even higher due to the lost cases through the calibration.

*Table 4* Logistic Regressions of the Probability of Missing Values (1 = yes) in Dependence of Design Features

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Style (ref. text) | -0.039<br>(0.391) | -0.165<br>(0.736) | -0.023<br>(0.562) | -0.042<br>(0.390) |
| Answering scale (ref. rating scale) | 1.104*<br>(0.432) | 1.019<br>(0.572) | 1.104*<br>(0.431) | 0.932<br>(0.592) |
| Order (ref. random order) | 0.042<br>(0.390) | 0.041<br>(0.390) | 0.058<br>(0.539) | -0.214<br>(0.737) |
| Style * answering scale | | 0.176<br>(0.867) | | |
| Style * order | | | -0.033<br>(0.780) | |
| Order * answering scale | | | | 0.359<br>(0.869) |
| Constant | -3.512***<br>(0.458) | -3.450***<br>(0.500) | -3.520***<br>(0.483) | -3.391***<br>(0.503) |
| McFaddens Pseudo $R^2$ | 0.034 | 0.034 | 0.034 | 0.035 |
| $N_{vignettes}$ | 8680 | 8680 | 8680 | 8680 |
| $N_{respondents}$ | 434 | 434 | 434 | 434 |

*Notes*. β-coefficients (log-odds) with cluster-robust (cluster=respondent) standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Proposed advantages of open scales are that they allow for more nuanced, fine-grained ratings of respondents. However, it is unclear if the respondents use the scale in the intended (metric) way. Table 5 shows the 10 most frequent values gained from the open scale. As it can be seen, respondents frequently used rough, rounded numbers (such as 100, 1000) to express their perception of injustice and did not fully exploit the open continuum of the scale.

Open scales are particularly deemed to perform better regarding the prevention of ceiling effects that could occur especially in a random order design. Table 6 provides the tests for differences in regression coefficients by vignette order separately for both scales. We use the multi-level linear regression models (shown in Model 1 and Model 3) and compare them to Tobit regressions that are regularly used to account for censored data (shown in Model 2 and Model 4). The joined test for differences across design features shows insignificant $\chi^2$-values for the linear models and insignificant F-values for the interactions specified via Tobit regression models. Thus, the more nuanced regression analyses correcting for a possible censoring of responses that are presented in Table 6 are in line with the more general results reported in Table 3, Model 3: Overall, the differences between the modes

*Table 5* Ten Most Frequent Values Indicated by Respondents on the Open Scale

| Value | *N* | Percent |
|---|---|---|
| 0 | 1282 | 36.57 |
| 100 | 319 | 9.10 |
| 10 | 201 | 5.73 |
| 1000 | 199 | 5.68 |
| 50 | 164 | 4.68 |
| 5 | 99 | 2.82 |
| 20 | 75 | 2.14 |
| 3 | 70 | 2.00 |
| 500 | 70 | 2.00 |
| 1 | 69 | 1.97 |

*Table 6* Tests for Vignette Order Effects on Vignette Evaluations

| | | Rating scale | | Open scale | |
|---|---|---|---|---|---|
| | | (1) Linear regression | (2) Tobit regression | (3) Linear regression | (4) Tobit regression |
| | df | $\chi^2$ | F | $\chi^2$ | F |
| Extreme cases first x sex | 1 | 0.024 | 0.002 | 0.059 | 0.106 |
| Extreme cases first x age | 1 | 0.374 | 0.184 | 2.937 | 1.044 |
| Extreme cases first x degree | 2 | 6.404* | 2.849 | 8.887* | 1.792 |
| Extreme cases first x children | 1 | 4.075* | 3.256 | 0.591 | 0.576 |
| Extreme cases first x experience | 1 | 0.002 | 0.001 | 0.186 | 0.161 |
| Extreme cases first x tenure | 1 | 0.821 | 1.247 | 0.871 | 2.998 |
| Extreme cases first x earnings | 1 | 1.255 | 1.149 | 2.099 | 1.799 |
| Extreme cases first x occupation | 9 | 11.451 | 1.252 | 8.385 | 0.919 |
| Overall | 17 | 22.735 | 1.348 | 21.116 | 1.016 |
| $N_{vignettes}$ | | 4389 | 4389 | 3506 | 3506 |
| $N_{respondents}$ | | 222 | 222 | 186 | 186 |

*Notes*. Tests after multi-level estimation with interaction terms; df: degrees of freedom; reference category: random order; Sidak-adjusted *p*-values; * $p < .05$; ** $p < .01$; *** $p < .001$.

of sorting are marginal for both types of answering scales. A closer look on the coefficients shows that with the rating scale there are two vignette dimensions (educational degree and children) that significantly differ depending on the order of the vignettes (Model 1). In case of extreme-cases-first ordering, the coefficients of

these dimensions are bigger in absolute size compared to those in the mode of random order, indicating potential ceiling effects. However, we also find one significant difference (again, for educational degree) with the open scale (Model 3). This is, however, only one positive finding within 17 tests. Performing Tobit regressions to account for ceiling effects (with cluster-robust standard errors accounting for the nested data structure) completely vanishes the significant differences between the experimental splits (Models 2 and 4).

In a final step, the experimental splits are evaluated regarding response times and response consistency (based on the squared residuals, see Table 7). Model 1 shows the results of a median regression of response time on the design features. The constant indicates that on the average, respondents needed about 17 seconds to evaluate a single vignette. While there were no differences for table vs. text vignettes and for different order, the use of open answering scale took on average about 3.5 seconds longer than the rating scale. This seems obvious since the evaluation using the open scale is based on a three-step process. A more nuanced picture of the response time by vignette position offers Figure 1 and shows a well-known pattern. Respondents need more time during the first vignettes in all experimental splits to get used to the task. They speed up until the fourth vignette and have a roughly stable response time then. When comparing different modes, it becomes obvious that the respondents using the open scale need always some seconds more due to the more complex rating task. Besides this difference, the patterns are similar in all experimental splits. The analysis of response consistency shown in Model 2 of Table 7 highlights differences between the answering scales with open scales producing higher squared residuals. We find no differences between other design features and also no interaction effects between design features (not shown).

*Table 7* Response Time and Response Consistency (Squared Residuals) by Experimental Variation

| | (1) Response time | (2) Residuals sq. |
|---|---|---|
| Style (ref. text) | -0.984 | -0.0593 |
| | (0.593) | (0.167) |
| Answering scale (ref. rating scale) | 3.531*** | 0.556** |
| | (0.618) | (0.185) |
| Order (ref. random order) | 0.312 | -0.0689 |
| | (0.589) | (0.168) |
| Constant | 17.12*** | 0.527* |
| | (0.989) | (0.244) |
| N | 7895 | 7895 |
| N_respondents | 408 | 408 |

*Note*: Coefficients of Model 1 are based on a median regression with cluster robust standard errors. Coefficients of Model 2 are based on a multi-level regression (GLS) with robust standard errors. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

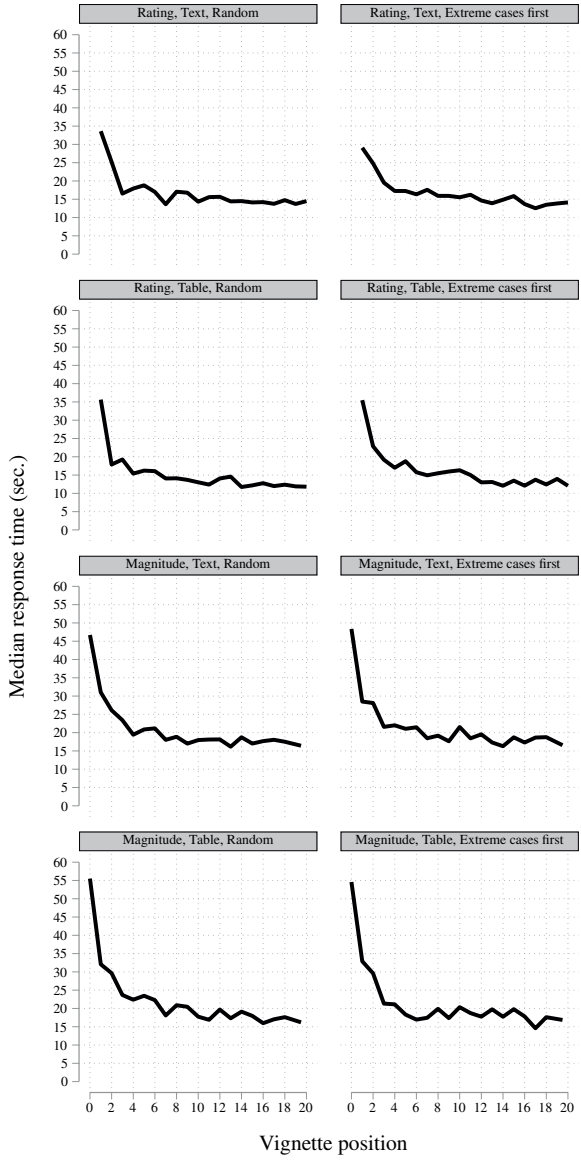


*Figure 1* Median response time in seconds per experimental variation (rating vs. magnitude answering scale, text vs. table, random order vs. extreme cases first) and vignette position. Note, in the vignettes with the open (magnitude) scale every respondent rated the same vignette (vignette position = 0) before the deck with 20 vignettes started. Therefore, the figures for the rating task start at vignette position 1 and the others at vignette position 0.

## Summary

This study analyzed the effects of design features of factorial surveys that have not been systematically evaluated so far, although these features are often varied across applications. We summarize the main findings in three implications and recommendations:

1. The presentation of dimensions in a running text – as it is done in most factorial surveys – did not produce significantly different results compared to a presentation in a table format. Our findings are in line with the study of Shamon et al. (2019) that also finds no differences between texts and tables focusing on response inconsistency and response time. However, their study finds differences between the two styles regarding the prevalence of missing values while we do not find differences. Shamon et al. (2019) find significantly lower total non-response (including refusals, break-offs, and vignette non-response) for table vignettes compared to text vignettes. They report about 24.1 percent of missing values for the vignette evaluations with most of them (18.3 percent) occurring due to refusals (i.e. respondents produced only missings in the vignette module or answered with a constant rating pattern). Focusing only on vignette non-response (without refusals) they report similar non-response numbers as we have (about 3.5 percent) and find support for text vignettes compared to table vignettes (less missing values). In our study we have only vignette non-response (2.9 percent with rating scales and 8.1 percent with magnitude scales) as nobody refused to fulfill the task. One explanation for different findings might be the different sample populations in both studies (in our study university students vs. quota sample of German population in Shamon et al. 2019) as well as the survey mode. We would expect that this difference is related both to the difference in population and survey mode, as well as the difference in the evaluation task. Taken together, we conclude that researchers might use tables instead of running texts, specifically if they want to neutralize possible effects of dimension order (see Auspurg & Jäckle, 2017), as tables allow for a more flexible (random) ordering of dimensions.
2. The rating scale clearly out-performed the open scale in many terms, e.g., in the number of missing values, and probably also produced more valid regression estimates. The open scales are more time consuming as a thorough introduction into the procedure and a calibration vignette is needed and, in our case, a three-step scale was necessary. In addition, the open scales did not come with the benefits of true metric scales. The findings are in line with other research indicating weak performance of metric scales with extensive response options (Sauer et al., 2014). We therefore recommend using standard, one-step rating scales. As we compared rating scales to three-step open scales, future research

should investigate potential differences between rating scales and one-step open scales used by Shamon et al. (2019).

3. The variation of the vignette order (random vs extreme-cases-first) did not yield to substantive differences in the overall estimation of regression coefficients. Only when splitting the analysis additionally by response scales, results slightly differed. Given these small differences, the easier and more flexible random sorting of vignettes seems quite more advisable. In case there occur ceiling effects, these can still be adjusted by means of specific econometric regression methods (cf. Auspurg & Hinz, 2015). Moreover, if ceiling effects occur in pre-tests, one might lessen them by switching to a broader rating scale (e.g., 11 points instead of 7) or lower numbers of vignettes.

# Conclusions

Our study found only few method effects, which is good news: Factorial survey results seemed to be very robust against the tested variations of design features. However, an exception existed with open (magnitude) scales, which performed on many parameters worse than standard rating scales. Given the relatively common usage (and recommendation) of these response scales, this is an important finding. In standard survey research, these response scales were already abandoned due to similar problems as the ones found in our study (see, e.g., Schaeffer & Bradburn, 1989). However, in multi-factorial survey designs they have been still used until today to prevent censored responses. The latter were, however, hardly spotted in our survey. This makes us even more confident in our recommendation that also in multi-factorial survey experiments one should in future better rely on standard rating scales.

Our study also has limitations. The most important one is certainly that the participants were throughout university students. This standardization enabled us to have more power to detect pure effects of design features. But this specific population also impacts the generalizability of our findings to other samples, as this population is particularly used to read and process complex information (provided in tables). Thus, additional research with general population surveys is needed. In addition, one should test applications that are more prone to social desirability bias. Therewith, one could explore whether the evident presentation of dimensions in tables triggers more socially desirable evaluations as when potentially sensitive dimensions are embedded in a short story. Finally, we only tested one variant of open response scales that was bound to a three-step response procedure, and one specific survey mode (an online survey). Evaluations of other design variants are certainly desirable although they occur less likely in practice as we tested the most common designs.

In sum: The study shows that multi-factorial survey designs are robust against variations in presentation style and kind of vignette order but answering scales should be selected carefully.

## References


Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology, 6*(3), 128-138.

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments* (Vol. 175). Los Angeles: Sage Publications.

Auspurg, K., Hinz, T., & Liebig, S. (2009). *Complexity, learning effects, and plausibility of vignettes in factorial surveys.* Paper presented at the 104th Annual Meeting of the American Sociological Association, San Francisco.

Auspurg, K., Hinz, T., Liebig, S., & Sauer, C. (2015). The Factorial Survey as a Method for Measuring Sensitive Issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, & P. Sturgis (Eds.), *Improving Survey Methods. Lessons from recent Research* (pp. 137-149). New York: Routledge.

Auspurg, K., & Jäckle, A. (2017). First equals most important? Order effects in vignette-based measurement. *Sociological Methods & Research, 46*(3), 490-539.

Dülmer, H. (2007). Experimental Plans in Factorial Surveys: Random or Quota Design? *Sociological Methods & Research, 35*(3), 382-409.

Dülmer, H. (2016). The Factorial Survey Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods & Research, 45*(2), 304-347.

Garret, K. (1982). Child abuse: Problems of definition. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments. The factorial survey approach* (pp. 177-204). Beverly Hills: Sage.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences, 112*(8), 2395-2400.

Hox, J. J., Kreft, I. G., & Hermkens, P. L. (1991). The analysis of factorial surveys. *Sociological Methods & Research, 19*(4), 493-510.

Jann, B. (2014). Plotting regression coefficients and other estimates. *Stata Journal, 14*(4), 708-737.

Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research, 34*(3), 334-423.

Jasso, G., & Rossi, P. H. (1977). Distributive justice and earned income. *American Sociological Review, 42*(4), 639-651.

Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research, 31*(4), 545-557.

Liebig, S., Sauer, C., & Friedhoff, S. (2015). Using factorial surveys to study justice perceptions: five methodological problems of attitudinal justice research. *Social Justice Research, 28*(4), 415-434.

O'Toole, R., Webster, S. W., O'Toole, A. W., & Lucal, B. (1999). Teachers' recognition and reporting of child abuse: a factorial survey. *Child Abuse and Neglect, 23*(11), 1083-1101.

Parente, P. M. D. C., & Santos Silva, J. M. C. (2016). Quantile Regression with Clustered Data. *Journal of Econometric Methods, 5*, 1-15.

Petzold, K., & Wolbring, T. (2019). What can we learn from factorial surveys about human behavior? *Methodology, 15*(1), 19-30.

Sauer, C., Auspurg, K., Hinz, T., & Liebig, S. (2011). The application of factorial surveys in general population samples: The effects of respondent age and education on response times and response consistency. *Survey Research Methods, 5*(3), 89-102.

Sauer, C., Auspurg, K., Hinz, T., Liebig, S., & Schupp, J. (2014). *Method effects in factorial surveys: An analysis of respondents' comments, interviewers' assessments, and response behavior.* SOEP papers on Muldisciplinary Panel Research, No. 629/2014. German Socio-Economic Panel Study (SOEP). Berlin.

Schaeffer, N. C., & Bradburn, N. M. (1989). Respondent behavior in magnitude estimation. *Journal of the American Statistical Association, 84*(406), 402-413.

Shamon, H., Dülmer, H., & Giza, A. (2019). The Factorial Survey: The Impact of the Presentation Format of Vignettes on Answer Behavior and Processing Time. *Sociological Methods & Research*, doi: 0049124119852382.

Shepelak, N. J., & Alwin, D. F. (1986). Beliefs about inequality and perceptions of distributive justice. *American Sociological Review, 51*(1), 30-46.

StataCorp. (2013). *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual neural and social prospects*. New York: Wiley.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research, 38*(3), 505-520.

Weinberg, J. D., Freese, J., & McElhattan, D. (2014). Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample. *Sociological Science, 1*(19), 292-310.

Wooldridge, J. M. (2003). *Introductory Econometrics. A Modern Approach*. Mason, OH: South Western.

methods, data, analyses | Vol. 14(2), 2020, pp. 215-240 DOI: 10.12758/mda.2020.03

# Combining Quantitative Experimental Data with Web Probing: The Case of Individual Solutions for the Division of Labor Between Both Genders

*Michael Braun*[1], *Katharina Meitinger*[2] & *Dorothée Behr*[1]

[1] *GESIS – Leibniz Institute for the Social Sciences, Germany*

[2] *Utrecht University, Netherlands*

**Abstract**

In 2012, a new question was introduced into the International Social Survey Program (ISSP). It asks respondents to indicate what they consider the best division of labor between men and women. In this paper, we propose to assess the validity and cross-national comparability of this new ISSP question, using a mixed-methods approach that combines quantitative experimental data with qualitative probing data. We implemented our experiment in non-probability online surveys in five countries, in which half of the respondents received the original ISSP question and the other half a variant with an additional category saying "Each family should find the solution which works best for them." In addition, the understanding of "individual solutions" was probed. We report on the understanding of this category.

*Keywords*: Web probing, cross-cultural research, mixed methods, gender roles, ISSP

Since 1985, the International Social Survey Program (ISSP) has conducted studies on different areas of social science research and thereby produced a huge data base for comparisons across countries and time. The majority of questions and items are held constant and kept unchanged over the different replications, but some questions are replaced in order to improve measurement quality or capture new trends. One of the topical modules of the ISSP is on "Family and Changing Gender Roles", which was fielded in 1988, 1994, 2002, and 2012. The way gender ideology was measured in the earlier surveys has often been criticized for having a traditional slant, focusing exclusively on women and employment, or for having methodological problems (Braun, 2008; Edlund & Öun, 2016). Though, from early on, there have also been attempts by researchers to construct more differentiated instruments that partly also capture subtle sexism (Brogan & Kutner, 1976; Glick & Fiske, 1997; King & King, 1997; Swim et al., 1995), these measure have not been adopted by large scale-comparative surveys.

In order to improve the measurement in the ISSP, in the 2012 round, a new measure for gender ideology was included to address respondents' preferences for the division of labor between men and women when there are children at home (ISSP Research Group, 2016; Scholz et al., 2014). Six types of preferences were presented as response categories, ranging from *the mother stays at home and the father works full-time* to the opposite division of labor (see further down for more details). Respondents should indicate what, according to their opinion, was the best way to organize the division of labor for a couple. This question forced respondents to single out one specific division of labor between men and women. Such a choice could be difficult for respondents who think that the best solution should be made dependent on additional considerations. For example, some respondents might think that the best solution should depend on the preferences of the partners, their abilities or their earning potential. Such respondents might struggle to choose one of the categories offered to them, and this might encourage superficial and stereotypical answer behavior. Therefore, when designing the items, the ISSP drafting group discussed whether an additional "individual solutions" category should be added. If so, this would give respondents who do not find their preferences represented in the answer categories an appropriate way out without having to either opt for "don't know" or select one of the substantive categories they do not really approve. The addition of such an "individual solutions" category, however, was eventually declined on the basis of concerns that in particular traditional respondents might use this category in order to avoid an overt disclosure of their traditional stance due to social-desirability considerations.

*Direct correspondence to*
Michael Braun, GESIS – Leibniz Institute for the Social Sciences, Germany
E-mail: michael.braun@gesis.org

As to the new item in the ISSP, this should be thoroughly assessed and checked for measurement equivalence across countries before it is used in substantive research. The most commonly used statistical technique for assessing measurement equivalence is multiple-group confirmatory factor analysis (MGCFA, Jöreskog, 1971). Latent-class analysis also has a long tradition in this field (Clogg, 1984; for an application to gender-role items of the World Value Survey and the European Values Study, see Knight & Brinton, 2017). Other techniques include correspondence analysis (for an application to the ISSP gender role items, see Blasius & Thiessen, 2006). All these quantitative methods are helpful in deciding whether measures are equivalent across countries but they usually do not allow getting at the causes of non-equivalence.[1] Much can be gained from getting at the causes of non-equivalence as well as from understanding the interpretations of respondents from different countries. Such interpretation patterns can be used in substantive research to avoid wrong conclusions. In addition, these quantitative methods cannot be applied to single items but to multiple-item measures only. Thus, they cannot be used to assess the new ISSP item on the division of labor between men and women. This is where qualitative approaches can and should come in.

Qualitative approaches, in particular cognitive interviews, are helpful to investigate problems in the response process (Beatty & Willis, 2007; Willis, 2005). A variety of probing techniques exist that are used during cognitive interviewing. For example, category-selection probes help to reveal the reasons for the selection of the responses to closed questions ("Please explain why you selected 'strongly agree'"). Unfortunately, international comparative cognitive studies drastically increase the coordination effort and are quite time-consuming (Willis, 2015) and, thus, are not implemented frequently in research (for exceptions see: Benítez et al., 2018; Fitzgerald et al., 2011; Miller et al., 2011; Thrasher et al., 2011; for a review see Willis, 2015).

However, the conduct of additional web-based studies to capture cross-cultural qualitative information is a potential source of information. "Web probing, that is, the implementation of probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey items" (Behr et al., 2017), is a method to complement quantitative techniques to establish measurement equivalence of items in cross-cultural research (Behr et al., 2017; Meitinger, 2017). In contrast to quantitative approaches that usually presuppose multiple-item measures, cognitive interviewing and web probing can also assess the cross-national comparability of single questions or items. In web probing studies, probing questions can be included in a regular web questionnaire. Behr & Braun (2015), for example,

1 While some quantitative approaches, such as multilevel structural equation modeling (MLSEM), can explain noninvariance by introducing macro-level variables in a multilevel analysis (Davidov et al., 2012), they are very demanding (e.g. samples should exceed 50 countries, see Meuleman & Billiet, 2009).

use a "category-selection" probe for a single item on satisfaction with democracy in order to find out which dimensions of democracy this question measures. The authors found that policy outcomes, governance, and aspects of the concrete political system play an important role in all countries of their study and, thus, answers can meaningfully be compared across countries.

Therefore, for assessing the consequences of including vs. excluding an "individual solutions" category, a mixed-methods approach seems to be particularly helpful (Creswell, 2014; Luyt, 2012; van de Vijver & Chasiotis, 2010). In the present case, we propose to combine the analysis of the quantitative survey data of the ISSP with a separate web study in which a split-half experiment with varying response categories was combined with a qualitative component. While the question experiment can inform the decision as to whether an "individual solutions" category matters in principal, the comparison of the web survey data with the data collected as part of the ISSP survey allows answering the question whether our results can be used to draw conclusions for the ISSP survey and its questionnaire.

# Data and Methods

## Sample

We implemented an experiment in non-probability online surveys in Germany, Great Britain, the United States, Mexico, and Spain with a total of 2,689 respondents. Survey participation was restricted to citizens of the respective countries aged 18 to 65. A net sample of approx. 500 respondents in each country was targeted using quotas for age (18-30, 31-50, and 51-65), gender, and education (lower vs. higher education). The panel providers were Respondi (www.respondi.com) and its partners in the respective countries. We met all quotas (see Table A1 in the Appendix for respective quota fields). Data collection was in June 2014. As these are quota samples, standardized response rates cannot be computed (Baker et al., 2010).

The selection of the five countries for the study was motivated by the expectation that in the liberal regime type (here represented by Great Britain and the United States) individuals or institutions outside of the family should not interfere with decisions regarding the roles of men and women in a family (compared to the conservative regime type here represented by Germany, Mexico, and Spain). These expectations should run in parallel to the lower involvement that the state has with regard to families (including the provision of a supporting infrastructure) in the first group of countries. Mexico was included alongside Germany and Spain as a strongly conservative country in which the family itself has a particularly high

importance in providing a support structure that might become relevant when it comes to the division of labor between both genders.

## Questionnaire

The International Social Survey Program (ISSP, ISSP Research Group 2016) asked the following new question in its 2012 "Family and Changing Gender Roles" module to capture respondents' views on the preferred division of labor between mother and father:

"Consider a family with a child under school age. What, in your opinion, is the best way for them to organize their family and work life?

1 The mother stays at home and the father works full-time.
2 The mother works part-time and the father works full-time.
3 Both the mother and the father work full-time.
4 Both the mother and the father work part-time.
5 The father works part-time and the mother works full-time.
6 The father stays at home and the mother works full-time."

In our web survey, half of the respondents received the original ISSP question (see Figure 1), the other half of the respondents received a variant (developed for this experiment) in which an additional category "Each family should find the solution which works best for them" was added. The respondents who selected the additional answer category also received a probing question regarding the reasons for opting for "individual solutions" (see Figure 2).

Thus, the experimental design combines quantitative insights from the split-ballot experiment with qualitative insights from web probing. To ensure the comparability of the probes themselves, we applied the team-driven TRAPD approach for the translation of the probes (Harkness, 2003).

**Consider a family with a child under school age.**
**What, in your opinion, is the best way for them to organise their family and work life?**

- ○ The mother stays at home and the father works full-time
- ○ The mother works part-time and the father works full-time
- ○ Both the mother and the father work full-time
- ○ Both the mother and the father work part-time
- ○ The father works part-time and the mother works full-time
- ○ The father stays at home and the mother works full-time

---

- ○ *can't choose*

*Figure 1* Experimental condition without response category "individual solutions"

**Consider a family with a child under school age.**
**What, in your opinion, is the best way for them to organise their family and work life?**

- ○ The mother stays at home and the father works full-time
- ○ The mother works part-time and the father works full-time
- ○ Both the mother and the father work full-time
- ○ Both the mother and the father work part-time
- ○ The father works part-time and the mother works full-time
- ○ The father stays at home and the mother works full-time
- ◉ Each family should find the solution which works best for them

---

- ○ *can't choose*

**Please explain why you selected "Each family should find the solution which works best for them ".**

The question was: "Consider a family with a child under school age. What, in your opinion, is the best way for them to organise their family and work life?"
- The mother stays at home and the father works full-time
- The mother works part-time and the father works full-time
- Both the mother and the father work full-time
- Both the mother and the father work part-time
- The father works part-time and the mother works full-time
- The father stays at home and the mother works full-time
- Each family should find the solution which works best for them
- *can't choose*

*Figure 2* Experimental condition with response category "individual solutions" and category-selection probe

## Translation of Open-ended Answers, Development of the Coding Scheme, and Coding

The Mexican and Spanish answers to the probe were translated into German by professional translators who had been briefed on the particularities of these texts as well as on translation and coding needs (Behr, 2015). The German and English answers were not translated but immediately coded by members of the project team (German native speakers with high proficiency in English).

An elaborated category scheme was developed, which represents the main criteria for the division of labor. This scheme was based on theory and also on the content of the probe responses.

Several theoretical perspectives can be found in the literature and based on these we developed hypotheses informing our probe scheme development. First, we wanted to investigate whether some of the approaches traditionally used to explain the actual household division of labor are also reflected in the reasoning of the respondents. These approaches are the *time-availability approach* that stipulates that spouses who spend more time working outside of the household show reduced participation with housework (Bianchi et al., 2000; Kalleberg & Rosenfeld, 1990) and the *resource-dependency approach* (Bittman et al., 2003; Brines, 1994) which recurs on the bargaining power of the spouses (based e.g. on their income or education) and its use to avoid unwanted housework. Second, we expected respondents to refer to individual preferences and capabilities, that is, what spouses want to do and where they are good at. Third, we surmised that several respondents would not recommend specific role distributions because they think that such decisions are the responsibility of the respective families or depend on the family's financial situation or on how child care can be organized (e.g. the presence of one parent or relatives at home or other alternative childcare arrangements).

The category scheme will be presented further below together with the results. Multiple coding was possible for all categories except for the categories *no generalization possible*, the *substantive rest category*, and *probe nonresponse*.

After the establishment of the final coding by members of the research team, a research assistant not involved in the development and implementation of the coding scheme coded 90% of the probe answers of all countries (while the other 10% were used for training purposes). Inter-rater agreement (between the final coding by members of the research team on the one hand and the research assistant on the other hand) ranged from 96% in Spain to 100% in the United States and Mexico. The high reliability value is likely to be a consequence of the relatively simple coding scheme, both as far as the number and the definitional clarity of the categories is concerned. This means that in more than 9 out of 10 cases, the raters coded a probing answer identically. All discrepancies of coding were discussed in the research team, which then arrived at a final version used in this paper.

## Analytical Strategy

In the following, we first compare the response pattern found in the ISSP data with the pattern revealed by our web survey to assess the general usefulness of our web survey data. Second, we compare the two experimental conditions implemented in the web survey. The first experimental condition asks the question on the best division of labor between father and mother exactly as it was in the ISSP, and the second experimental condition adds the answer category "Each family should find the solution which works best for them". Third, we report the responses to the category-selection probe regarding which "individual solutions" respondents had in mind when answering the closed question.

# Results

## Replication of the Pattern in the ISSP Data

A comparison of the first split of our web survey (which exactly replicates the original ISSP question) with the ISSP data[2] reveals that the general response pattern is replicated in our web survey (see Tables A2 and A3 in the Appendix). The ISSP and the web samples share the nearly complete lack of support for a role reversal and similar percentages of respondents who opt for the "don't know" category. In all five countries, the overwhelming majority supports the strict (only the father goes out to work) or moderate variant (the mother has only a complementary work role) of the male-breadwinner model if they are forced to choose among the models presented.

However, the respondents of the web survey seem to be less traditional than the respondents in the ISSP, despite of the quotas we have implemented for age, gender, and education.

Nevertheless, because of the experimental approach taken here, we are confident that the results found on the basis of the probing study can shed light on the ISSP data.

## "Individual Solutions" for the Division of Labor Between Both Genders

Table 1 shows the response distribution of the closed item in the web survey (the preferred division of labor between men and women), where the second split of the web survey contains the additional answer category "Each family should find

2 We did not restrict the ISSP data to the age range of the web survey.

*Table 1* Preferred division of labor dependent on the presence of an "individual solutions" category in the different countries (in percent)

| | Germany | | Great Britain | | United States | | Mexico | | Spain | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 |
| Mother at home, father full-time | 20 | 11 | 27 | 14 | 27 | 12 | 22 | 16 | 4 | 3 |
| Mother part-time, father full-time | 35 | 22 | 28 | 10 | 23 | 11 | 50 | 29 | 21 | 10 |
| Both full-time | 10 | 4 | 10 | 5 | 23 | 10 | 11 | 8 | 22 | 6 |
| Both part-time | 23 | 6 | 13 | 4 | 6 | 2 | 15 | 4 | 45 | 15 |
| Father part-time or at home, mother full-time | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| Individual solutions | - | 55 | - | 65 | - | 61 | - | 40 | - | 66 |
| Don't know | 13 | 1 | 20 | 2 | 20 | 3 | 2 | 1 | 8 | 0 |
| *N* | 275 | 264 | 281 | 253 | 266 | 274 | 253 | 292 | 268 | 263 |

*Data source*: Web survey; split 1: original ISSP version, split 2: "individual solutions" category added; original categories "father part-time and mother full-time" and "father at home and mother full-time" collapsed.

the solution which works best for them." When this individual-solutions category is introduced, clearly more than half of the respondents choose this category, with the only exception of Mexico (40%). In addition, when this answer category is provided, the prevalence of "don't know" responses drops drastically (from 2-20% to 0-3%). All other divisions of labor are chosen considerably less in the second split (with "individual solutions") compared to the first split (without "individual solutions"). However, the relative decrease is most marked for the "both part-time" category in most of the countries.

Though our experiment represents a between- instead of a within-subjects design, it seems nevertheless fair to conclude that the individual solutions category contains those respondents who would opt for the "don't know" option when the "individual solution" option is not available. In addition, the individual solutions category draws from all substantive categories and, in particular, from the "both part-time" response category. Part of the respondents choosing this category seem to use it as a compromise since none of the categories offered match their real preferences. Thus, it is this – not particularly traditional – category which loses support once the individual solutions category is added and not the more traditional answer

categories, as feared among questionnaire developers when designing the new ISSP question.

This can be seen even clearer from Table A4 in the Appendix which shows the web survey results if only the substantive ISSP categories (that is, without "don't know" and "individual solutions") are included in calculating the percentages. In all countries but the United States, it is the most traditional answer category that gains relative importance if an "individual solution" category is added (in the United States, it simply makes no difference). This applies to both genders (see Tables A5 and A6 in the Appendix). Table A7 in the Appendix shows the popularity of the individual solutions category in different social groups, in addition to gender. In most countries, those who opt for the individual solutions category are older than those who do not. Those who are married are less in favor of individual solutions than unmarried respondents. However, there are no consistent relationships between the choice of the individual solutions category and respondents' employment status and their partners' employment status and whether they have children or not.

Therefore, it is fair to conclude that adding an individual solutions category is not mainly used as an easy escape by traditional respondents who do not want to disclose their position in an overt manner. It might rather be used by those respondents who think that it impossible to opt for only one of the presented divisions of labor, unless more details on the specific situation of the respective family are taken into account.

## "Individual Solutions" Respondents have in Mind

What, then, are these "individual solutions"? Are respondents simply too lazy to make their choice among the answer categories offered or do they have concrete ideas in mind? This was the research goal we pursued with our open-ended probing question. Table 2 presents by country the types of "individual solutions" that respondents think of regarding the division of labor between men and women.

The first two codes that we extracted from the open-ended answers offered by our respondents, *time availability* and *resource dependency*, refer to general rules which depend less on personal decisions and preferences of the family or the partners involved. *Time availability* connects the decision on household labor to the labor-force involvement. Respondents refer to the time resources of both partners. The division of household labor should take into consideration how much time is left after paid work (e.g. "It depends on the jobs the parents have, whether it is possible to work part-time"). This argumentation pattern is gender neutral. It also leaves – as a general rule – open, how the division of market labor is established.

*Resource dependency* is broader in that it also connects the decision about who might work outside of the home and who might stay at home to the earning

*Table 2* Answers to the category-selection probe for respondents who opted for the "individual solutions" category in the closed question in the different countries (in percent)

| | Germany | Great Britain | United States | Mexico | Spain |
|---|---|---|---|---|---|
| Time availability | 10 | 2 | 3 | 12 | 17 |
| Resource dependency | 21 | 16 | 8 | 11 | 13 |
| Individual preferences | 8 | 7 | 4 | 3 | 5 |
| Individual abilities | 0 | 8 | 2 | 5 | 3 |
| Family/partners have to decide | | | | | |
| - no interference | 3 | 10 | 11 | 1 | 1 |
| - joint decision | 3 | 7 | 11 | 13 | 8 |
| - general | 10 | 5 | 6 | 2 | 1 |
| Situation dependency | | | | | |
| - financial necessities | 14 | 9 | 13 | 15 | 17 |
| - presence of one parent at home | 8 | 9 | 8 | 6 | 6 |
| Alternative possibilities | 13 | 5 | 7 | 10 | 17 |
| No generalization possible | 26 | 38 | 29 | 34 | 33 |
| Substantive rest category | 6 | 7 | 15 | 12 | 6 |
| Probe nonresponse | 3 | 2 | 1 | 0 | 1 |
| N | 144 | 164 | 168 | 117 | 173 |

*Data source*: Web survey, Split 2; multiple coding possible for all categories except for no generalization possible, substantive rest category, and probe nonresponse; that is, figures do not add up to 100%.

potential of the partners. The person with the higher earning potential or career opportunity should work outside the home. As the citation "Well it could be the case where the mother could earn more income in her job than the father could and therefore it would be better for the mother to work than the father" reveals, this argumentation pattern is – in principle – again gender neutral. Admittedly, this argumentation pattern – and the same applies to *time availability* – can be used by traditional respondents, too, especially when they surmise that men will earn more than women anyhow in most cases. For time availability an additional caveat is necessary if the amount of labor-force participation of the woman is not reflecting her free will but has been kept low by the intervention of the man. As a consequence, it is not possible to unambiguously gauge the traditionality of respondents who opt for these categories.

*Time availability* is a frequent criterion in Spain (17%) and, to a smaller degree also in Mexico (12%) and Germany (10%), but it is rarely used in Great Britain and the United States (2% and 3%, respectively). *Resource dependency* as a criterion is clearly more popular than time availability in Germany (21%) and the two Anglo-Saxon countries (16% in Great Britain and 8% in the United States), and of nearly equal importance as time availability in Mexico and Spain (11% and 13%, respectively).

The code *individual preferences* captures when respondents refer to the partners' interests and preferences which should decide on the division of labor ("Because some women and men would rather stay home and take care of their house or their kids and some want to work").

Country differences with regard to *individual preferences* are not pronounced, ranging from 3% in Mexico to 8% in Germany. In general, both of these codes are of minor importance compared to *time availability* and *resource dependency.*

The code *individual abilities* reflects capabilities of the partners with regard to the job and household chores or childraising as the main decision criterion (e.g. "Every home situation is personal. It depends on which parent has the best career prospects and ability to support the family but also who would be the most suitable parent to take more responsibility raising the children"). *Individual abilities* have a similar importance as *individual preferences* in most countries, ranging from 0% to 8%. However, what is striking is the complete absence of this criterion in Germany.

A further important criterion for the decision on individual solutions is the idea that the *family/partners have to decide by themselves*. Respondents differ in their focus: *No interference* stresses that the society or other people in general have no right to intervene in this private decision (e.g. "Democracy allows individual freedom. The State has no place interfering in personal lives"). *Emphasis on joint family* decisions indicates that a consensus in the family should be reached which might involve engaging in compromises (e.g. "… if it is agreeable to both parents"). Respondents also made rather general statements, which are not pronounced enough to be classified into one of the two previous codes (e.g. "You cannot offer a solution for all. That has to be individually decided by the respective families").

Overall, in Great Britain and the United States, respondents are clearly more in favor of the family or the partners to decide on the division of household labor than in the other countries. Both countries belong to the liberal regime type where the state is not assumed to intervene in family life and does neither actively facilitate nor hinder the combination of family and work roles (by men and women). This kind of individualism is expressed with most vigor in the *no interference* category which holds any outside intervention (and maybe even advice) into family decisions to be illegitimate: 10% of the respondents in Great Britain and 11% in the United States share this stance compared to only 1-3% in the other three countries.

Respondents also took the situational context into account when responding to the probe. *Situation dependency – financial necessities* applies when the organization of the role division should be decided taking the financial necessities of a family into account, in particular whether a double income is needed to make ends meet (e.g. "Sometimes it is necessary for both parents to work in order to financially provide for their child and family. However, if it is possible to live comfortably with just one parent working, then it is up to the parents to decide how they want to raise their family"). *Financial necessities* come to mind quite frequently. Spanish respondents think of this aspect most often (17%). This does not come as a surprise, as Spain is one of the countries which were most severely hit by the financial crisis (beginning in 2007) and the web survey was conducted during its peak/aftermath in 2014. On the contrary, British respondents are the least frequent to mention this aspect (9%).

In contrast, we assigned the code *situation dependency – presence of one parent at home*, when respondents favor a model in which one person goes out to work and the other cares for the children and the household. Whether the man or the woman goes out to work or stays home is irrelevant – at least to most respondents (e.g. "I believe that pre-school children benefit most from having a parent care for them full-time, but it does not matter if it is father or mother"). The call *for the presence of one parent at home* is of moderate frequency and country differences are relatively small, ranging from 6% in Mexico and Spain to 9% in Great Britain.

Respondents also thought of *alternative possibilities* to fulfill the needs of the children that are not related to the allocation of work roles among the parents. Examples are the involvement of grandparents as well as privately or publicly organized daycare (e.g. "There are various support systems available within different families, so no particular hard rule can apply in all instances").

Spanish and German respondents (17% and 13%, respectively) more often think of *alternative possibilities* (such as the involvement of grandparents or daycare) while in the Anglo-Saxon countries such a response is less frequent (5% in Great Britain and 7% in the United States). Contrary to our expectations, Mexicans are in-between.

*No generalization possible* was coded when respondents referred to individual differences in general without specifying any concrete criteria for the division of labor between both genders (e.g. "There is no right or wrong way, no one solution can suit every family"). Between one fourth (Germany) and more than one-third (Great Britain) of the respondents are coded into this category, thus referring to individual differences in general without specifying any concrete criteria. This could be an effect of web probing. Due to the web implementation, there is no possibility to spontaneously follow-up on answers that are not yet sufficiently clear.

The *substantive rest category* comprises answers that cannot be categorized into the substantive codes (e.g. "That again is freedom") or are difficult to compre-

hend. Between 6% of the respondents in Germany and Spain and 15% in the United States give a response that we coded as *substantive rest category*. This means that we could not fit the response into our category scheme (and similar responses did not occur frequently enough to justify the addition of additional categories) or it was not sufficiently comprehensible to assign it unambiguously to one of the existing codes. This is unfortunately a weakness of web probing, namely, that it does not allow for a clarification of unclear statements made by respondents (a problem that could be easily solved by the interviewer in a cognitive interview; see Meitinger & Behr, 2016).

Finally, *probe nonresponse* includes explicit refusals, "don't knows", and answers such as "dddfff". This topic, however, is not affected by *probe nonresponse*; nearly all respondents try to give a substantive answer.

In addition to the general prevalence of response categories, we also conducted an analysis of gender differences with regard to these codes. However, there are hardly any consistent differences between men and women across all countries (see Table A8 in the Appendix). In most countries, however, women are more likely to refer to *situation dependency – financial necessities* and *alternative possibilities* than men. On the contrary, they are less likely to opt for the *no generalization possible* category than men.

## Discussion

In this paper, we demonstrated the usefulness of web probing when there is only one item to validate. We used the example of a new instrument in the ISSP module on "Family and Changing Gender Roles" (2012). The new instrument asks respondents to select one out of six role divisions between men and women when there are children at home. An "individual solutions" category was not added in the original ISSP questionnaire due to some concern that traditional respondents might use this category to avoid an overt disclosure of their traditional positions.

Our results show, however, that the "individual solutions" category is likely to be used by all kinds of respondents, not only the traditional ones. This was revealed both by the experimental quantitative and the qualitative data. The experiment showed that the addition of an "individual solutions" category to the response alternatives of the ISSP question was most attractive for less traditional respondents who would otherwise opt for the "both part-time" response alternative, and to those who would otherwise choose the "don't know" category. The qualitative data from the web probing was likewise very informative, even though the single largest group of respondents in all countries referred to differences between individuals and families in a general way. Nevertheless, most respondents mentioned concrete criteria that should be used in families for coming to a decision on the optimal divi-

sion of labor between men and women. These criteria are mainly gender neutral, at least at face value. However, as mentioned above, as these criteria can also be used by traditional respondents (who take the inequality between both genders in the underlying conditions for granted), it is not possible to unambiguously infer the traditionality of respondents from these answers. Although the inclusion of an "individual solutions" category did not lead to the anticipated consequence (namely that it would attract mostly traditional respondents who did not want to explicitly express their position), it can nevertheless not be recommended for a regular survey that is not supplemented by web probing. This is because the selection of the "individual solutions" category cannot unambiguously be interpreted without the information from web probing and, thus, a clearly interpretable response would be missing for about half of the respondents.

Comparing the countries in our study, the majority of the criteria are of roughly the same importance in all countries or most of them. However, there are some noteworthy exceptions. In the two Anglo-Saxon countries (Great Britain and the United States), time availability was not mentioned frequently. In addition, in these two countries alternative possibilities of child care – outside the nuclear family – are less seen as a potential remedy to help decide on the role division between both partners. Instead, and in line with our hypothesis, in Great Britain and the United States, respondents make a point in that it is the family and the partners who have to decide this issue, and interference from outside of the family (in particular by the society at large) is seen as largely illegitimate.

In any case, the mixed-methods approach was crucial in assessing the consequences of adding an "individual solutions" category to the newly constructed ISSP item. We started with the quantitative ISSP data and compared it with the quantitative data of our web survey in order to establish whether it is possible to generalize results obtained from the latter to the former survey. Within the web survey, we then conducted a question experiment where the treatment group received an additional response category. Finally, this additional response category was probed and using the qualitative information obtained from the web survey the probe answers were coded and analyzed in a quantitative manner. The mixed-methods approach chosen allowed us to gain insights that we could not possibly have obtained by using a quantitative or qualitative method alone.

Several limitations of our study have to be mentioned. First, we used data based on non-probability online surveys. In order to tackle the issue whether we can use the web survey to shed light on the ISSP survey we compared the distribution of the central variable which was measured in the same way in one of the experimental splits and the ISSP survey. Nevertheless, we cannot exclude that results from probing could be somewhat different for the general population compared to the web survey.

Second, our experimental and probing data is limited to five countries and in these the highly developed countries are overrepresented compared to the ISSP survey. Only by replicating our study in additional countries in which the ISSP is conducted can we become more confident that our findings describe a general tendency in answer behavior and are not restricted to the countries we selected.

This paper does not inform on the more general question whether multiple-item measures are more adequate to measure gender-role attitudes than the single question we have analyzed. The evidence collected here is restricted to deciding in favor or against an inclusion of an "individual solution" category in the new ISSP question. In more general terms, while multi-item measures have clear advantages compared to a measure consisting of a single question (e.g. the possibility to employ data-analytic methods to establish equivalence across countries), there are also shortcomings with (existing) multi-items measures in large-scale comparative research. While most of the extant questions are concentrated on the role of the woman and might have a traditional slant, the construction of more balanced items which allow capturing egalitarian attitudes is also challenging, as there are a variety of possible egalitarian stances (Braun, 2008). In the end, the new ISSP measure was one attempt to bypass these shortcomings. At least in the area of gender-role attitudes, both question formats (multiple-item batteries and single questions) seem to have their merits (and weaknesses).

# References

Baker, R., Blumberg, S. J. , Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., … Lavrakas, P. J. (2010). AAPOR report on online panels. *Public Opinion Quarterly, 74*, 711–781.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*, 287–311.

Behr, D. (2015). Translating answers to open-ended survey questions in cross-cultural research: A case study on the interplay between translation, coding, and analysis. *Field Methods, 27*, 248–299.

Behr, D., & Braun, M. (2015). Satisfaction with the way democracy works: How respondents across countries understand the question. In P. B. Sztabinski, H. Domanski, & F. Sztabinski (Eds.), *Hopes and anxieties. Six waves of the European Social Survey (*pp. 121–138). Frankfurt am Main: Lang.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – Implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. Mannheim, GESIS – Leibniz Institute for the Social Sciences (GESIS – Survey Guidelines), DOI: 10.15465/gesis-sg_en_023.

Benítez, I., & Padilla, J.-L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research, 8*, 52-68.

Benítez, I., Padilla, J.-L., van de Vijver, F., & Cuevas, A. (2018). What cognitive interviews tell us about bias in cross-cultural research: An illustration using quality-of-life items. *Field Methods*, *30*, 277-294.

Bianchi, S. M., Milkie, M. A., Sayer, L. C., & Robinson, J. P. (2000). Is anyone doing the housework? Trends in the gender division of household labor. *Social Forces, 79*, 191-228.

Bittman, M., England, P., Sayer, L. C., Folbre, N., & Natheson, G. (2003). When does gender trump money? Bargaining the time in household work. *American Journal of Sociology, 109*, 186-214.

Blasius, J., & Thiessen, V. (2006). Assessing data quality and construct comparability in cross-national surveys. *European Sociological Review, 22*, 229-242.

Braun, M. (2008). Using egalitarian items to measure men's and women's family roles. *Sex Roles, 59*, 644-656.

Brines, J. (1994). Economic dependency, gender, and the division of labor at home. *American Journal of Sociology, 100*, 652-688.

Brogan, D., & Kutner, N.G. (1976). Measuring sex-role orientation: a normative approach. *Journal of Marriage and the Family, 38*, 31-40.

Clogg, C. C. (1984). Some statistical models for analyzing why surveys disagree. In Turner, C. F., & Martin, E. (Eds.), *Surveying subjective phenomena. Volume 2* (pp. 319–366). New York: Russell Sage.

Creswell, J. W. (2014). *Research design. Qualitative, quantitative, and mixed methods approaches.* 4th edition. Los Angeles: Sage.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P. & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology, 43*, 558-575.

Edlund, J., & Öun, I. (2016). Who should work and who should care? Attitudes towards the desirable division of labour between mothers and fathers in five European countries. *Acta Sociologica, 59*, 151-169.

Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2011). Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. *Journal of Official Statistics, 27*, 569–599.

Glick, P. & Fiske, S. T. (1997). Hostile and benevolent sexism. Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly, 21*, 119-135.

Harkness, J. A. (2003). Questionnaire translation. In Harkness, J. A., van de Vijver, F. J., & Mohler P. (Eds.), *Cross-cultural survey methods* (pp. 35-56). New York: Wiley.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

ISSP Research Group (2016). International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012. GESIS Data Archive, Cologne. ZA5900 Data file Version 4.0.0, DOI: 10.4232/1.12661.

Kalleberg, A. L. & Rosenfeld, R. A. (1990). Work in the family and in the labor market: A cross-national, reciprocal analysis. *Journal of Marriage and the Family, 52*, 331-346.

King, L. A., & King, D. W. (1997). Sex-role egalitarianism scale. Development, psychometric properties, and recommendations for future research. *Psychology of Women Quarterly, 21*, 71-87.

Knight, C. R., & Brinton, M. (2017). One egalitarianism or several? Two decades of gender-role attitude change in Europe. *American Journal of Sociology, 122*, 1485-1532.

Luyt, R. (2012). A framework for mixing methods in quantitative measurement development, validation, and revision: a case study. *Journal of Mixed Methods Research, 6*, 294–316.

Meitinger, K. 2017. Necessary but insufficient: Why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly, 81*, 447–472.

Meitinger, K. & Behr, D. (2016). Comparing cognitive interviewing and online probing: Do they find similar results? *Field Methods, 28*, 363-380.

Meuleman, B. & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods, 3*, 45-58.

Miller, K., Fitzgerald, R., Padilla, J.-L., Willson, S., Widdop, S., Caspar, R., … Schoua-Glusberg, A. (2011). Design and analysis of cognitive interviews for comparative multinational testing. *Field Methods, 23*, 379–396.

Scholz, E., Jutz, R., Edlund, J., Öun, I., & Braun, M. (2014). ISSP 2012. Family and Changing Gender Roles IV. *GESIS-Technical Reports 2014/19.* Köln: GESIS. https://www.gesis.org/issp/modules/issp-modules-by-topic/family-and-changing-gender-roles/2012/

Swim, J. K., Aikin, K. J., Hall, W. S, & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology, 68*, 199-214.

Thrasher, J. F., Quah, A. C., Dominick, G., Borland, R., Driezen, P., Awang, R., ... Boado, M. (2011). Using cognitive interviewing and behavioral coding to determine measurement equivalence across linguistic and cultural groups: An example from the International Tobacco Control Policy Evaluation Project. *Field Methods, 23*, 439–460.

van de Vijver, F. J. R., & Chasiotis, A. (2010). Making methods meet: Mixed designs in cross-cultural research. In Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., … Smith, T. (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 455–473). Hoboken, NJ: Wiley.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: Sage.

Willis, G. B. (2015). The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly, 79*, 359–395.

# Appendix

*Table A1* Quota assignment in the web survey

| Age | Gender | Education | Germany (539) | Great Britain (534) | United States (540) | Spain (531) | Mexico (545) |
|---|---|---|---|---|---|---|---|
| 18-30 | Male | High | 8.53 | 8.24 | 8.15 | 8.29 | 8.99 |
| 18-30 | Male | Low | 8.16 | 8.24 | 8.15 | 8.29 | 8.26 |
| 18-30 | Female | High | 8.91 | 8.61 | 8.15 | 8.29 | 8.62 |
| 18-30 | Female | Low | 8.35 | 8.80 | 8.52 | 8.29 | 8.26 |
| 31-50 | Male | High | 8.16 | 8.24 | 8.33 | 8.29 | 8.62 |
| 31-50 | Male | Low | 8.35 | 8.24 | 8.52 | 8.66 | 8.07 |
| 31-50 | Female | High | 8.16 | 8.24 | 8.15 | 8.29 | 8.26 |
| 31-50 | Female | Low | 8.35 | 8.24 | 8.52 | 8.29 | 8.07 |
| 51-65 | Male | High | 8.16 | 8.24 | 8.33 | 8.47 | 8.07 |
| 51-65 | Male | Low | 8.16 | 8.24 | 8.33 | 8.29 | 8.07 |
| 51-65 | Female | High | 8.53 | 8.33 | 8.70 | 8.29 | 8.44 |
| 51-65 | Female | Low | 8.16 | 8.24 | 8.15 | 8.29 | 8.26 |
| | | | 100% | 100% | 100% | 100% | 100% |

*Table A2* Preferred division of labor for ISSP question in the different countries (in percent)

| | Germany | | Great Britain | | United States | | Mexico | | Spain | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web |
| Mother at home, father full-time | 20 | 20 | 34 | 27 | 29 | 27 | 49 | 22 | 24 | 4 |
| Mother part-time, father full-time | 44 | 35 | 38 | 28 | 32 | 23 | 23 | 50 | 39 | 21 |
| Both full-time | 10 | 10 | 4 | 10 | 9 | 23 | 7 | 11 | 11 | 22 |
| Both part-time | 13 | 23 | 4 | 13 | 5 | 6 | 16 | 15 | 18 | 45 |
| Father part-time or at home, mother full-time | 1 | 1 | 0 | 1 | 1 | 1 | 3 | 0 | 1 | 0 |
| Don't know | 13 | 13 | 20 | 20 | 25 | 20 | 3 | 2 | 7 | 8 |
| N | 1,766 | 275 | 950 | 281 | 1,302 | 266 | 1,527 | 253 | 2,595 | 268 |

*Data source*: ISSP 2012; Web survey, split 1; original categories "father part-time and mother full-time" and "father at home and mother full-time" collapsed.

*Table A3* Preferred division of labor for ISSP question in the different countries, separately for male and female respondents (in percent)

| | Germany | | | | Great Britain | | | | United States | | | | Mexico | | | | Spain | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | | Female | | Male | | Female | | Male | | Female | | Male | | Female | | Male | | Female | |
| | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web | ISSP | Web |
| Mother at home, father full time | 24 | 26 | 17 | 14 | 38 | 27 | 30 | 27 | 34 | 29 | 25 | 26 | 48 | 26 | 50 | 17 | 29 | 4 | 20 | 4 |
| Mother part time, father full time | 41 | 29 | 47 | 41 | 35 | 30 | 41 | 26 | 30 | 20 | 33 | 26 | 23 | 53 | 22 | 46 | 37 | 18 | 41 | 24 |
| Both full time | 10 | 12 | 9 | 7 | 4 | 11 | 4 | 10 | 8 | 26 | 9 | 20 | 7 | 9 | 7 | 15 | 12 | 25 | 11 | 19 |
| Both part time | 11 | 21 | 14 | 24 | 4 | 11 | 5 | 15 | 3 | 6 | 6 | 6 | 16 | 11 | 15 | 20 | 15 | 43 | 20 | 48 |
| Father part time or at home, mother full time | 0 | 1 | 1 | 0 | 0 | 3 | 0 | 0 | 1 | 1 | 0 | 1 | 2 | 0 | 3 | 0 | 1 | 0 | 1 | 0 |
| Don't know | 14 | 12 | 12 | 15 | 20 | 18 | 20 | 22 | 24 | 18 | 26 | 21 | 4 | 1 | 3 | 2 | 7 | 11 | 7 | 6 |
| *N* | 857 | 145 | 909 | 130 | 438 | 152 | 512 | 129 | 594 | 129 | 708 | 137 | 727 | 138 | 796 | 115 | 1,271 | 134 | 1,378 | 134 |

*Data source:* ISSP 2012; Web survey, split 1.

*Table A4* Preferred division of labor dependent on the presence of an "individual-solutions" category in the different countries, separately for male and female respondents (in percent)

| | Germany | | | | Great Britain | | | | United States | | | | Mexico | | | | Spain | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | | Female | | Male | | Female | | Male | | Female | | Male | | Female | | Male | | Female | |
| | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 |
| Mother at home, father full time | 26 | 13 | 14 | 10 | 27 | 14 | 27 | 13 | 29 | 14 | 26 | 9 | 26 | 19 | 17 | 14 | 4 | 4 | 4 | 2 |
| Mother part time, father full time | 29 | 24 | 41 | 21 | 30 | 13 | 26 | 9 | 20 | 12 | 26 | 10 | 53 | 30 | 46 | 28 | 18 | 8 | 24 | 12 |
| Both full time | 12 | 6 | 7 | 3 | 11 | 7 | 10 | 3 | 26 | 14 | 20 | 6 | 9 | 10 | 15 | 7 | 25 | 6 | 19 | 7 |
| Both part time | 21 | 7 | 24 | 6 | 11 | 3 | 15 | 4 | 6 | 4 | 6 | 1 | 11 | 4 | 20 | 5 | 43 | 17 | 48 | 12 |
| Father part time or at home, mother full time | 1 | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| Individual solutions | - | 49 | - | 59 | - | 61 | - | 68 | - | 50 | - | 73 | - | 36 | - | 44 | - | 65 | - | 67 |
| Don't know | 12 | 2 | 15 | 1 | 18 | 2 | 22 | 3 | 18 | 5 | 21 | 1 | 1 | 1 | 2 | 1 | 11 | 0 | 6 | 0 |
| *N* | 145 | 122 | 130 | 142 | 152 | 112 | 129 | 141 | 129 | 140 | 137 | 134 | 138 | 135 | 115 | 157 | 134 | 133 | 134 | 130 |

*Data source:* Web survey; split 1: original ISSP version, split 2: "individual-solutions" category added.

*Table A5* Preferred division of labor dependent on the presence of an "individual-solutions" category in the different countries (in percent; calculation excluding "don't know" and "individual solutions" categories)

| | Germany | | Great Britain | | United States | | Mexico | | Spain | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 |
| Mother at home, father full-time | 23 | 26 | 34 | 42 | 34 | 33 | 22 | 27 | 4 | 9 |
| Mother part-time, father full-time | 40 | 50 | 35 | 31 | 29 | 32 | 51 | 49 | 23 | 28 |
| Both full-time | 11 | 9 | 13 | 15 | 29 | 28 | 12 | 14 | 24 | 19 |
| Both part-time | 26 | 15 | 16 | 11 | 8 | 6 | 15 | 8 | 49 | 43 |
| Father part-time or at home, mother full-time | 0 | 0 | 2 | 1 | 1 | 2 | 0 | 2 | 0 | 1 |
| *N* | 239 | 117 | 224 | 83 | 214 | 98 | 249 | 171 | 246 | 90 |

*Data source:* Web survey; split 1: original ISSP version, split 2: "individual-solutions" category added; original categories "father part-time and mother full-time" and "father at home and mother full-time" collapsed.

*Table A6* Preferred division of labor dependent on the presence of an "individual-solutions" category in the different countries, separately for male and female respondents (in percent; calculation excluding "don't know" and "individual solutions")

| | Germany | | | | Great Britain | | | | United States | | | | Mexico | | | | Spain | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | | Female | | Male | | Female | | Male | | Female | | Male | | Female | | Male | | Female | |
| | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 | Split 1 | Split 2 |
| Mother at home, father full time | 29 | 27 | 16 | 25 | 33 | 38 | 35 | 46 | 36 | 32 | 32 | 34 | 26 | 29 | 17 | 26 | 5 | 11 | 4 | 7 |
| Mother part time, father full time | 33 | 48 | 48 | 53 | 37 | 33 | 33 | 29 | 25 | 27 | 33 | 40 | 54 | 47 | 47 | 51 | 20 | 21 | 25 | 35 |
| Both full time | 13 | 12 | 8 | 7 | 13 | 19 | 13 | 10 | 31 | 30 | 26 | 23 | 9 | 15 | 15 | 13 | 28 | 17 | 20 | 21 |
| Both part time | 24 | 13 | 28 | 16 | 14 | 7 | 19 | 15 | 8 | 8 | 7 | 3 | 11 | 6 | 20 | 9 | 48 | 49 | 51 | 37 |
| Father part time or at home, mother full time | 1 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 1 | 3 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 2 | 0 | 0 |
| *N* | 128 | 60 | 111 | 57 | 124 | 42 | 100 | 41 | 106 | 63 | 108 | 35 | 136 | 85 | 113 | 86 | 120 | 47 | 126 | 43 |

*Data source:* Web survey; split 1: original ISSP version, split 2: "individual-solutions" category added.

*Table A7* Choice of individual-solution category in different social groups (in percent)

| | Germany | Great Britain | United States | Mexico | Spain |
|---|---|---|---|---|---|
| Women | 59 | 68 | 73 | 44 | 67 |
| Men | 49 | 61 | 50 | 36 | 65 |
| Average year of birth (individual-solutions category not selected) | 1974 | 1974 | 1973 | 1975 | 1976 |
| Average year of birth (individual-solutions category selected) | 1971 | 1970 | 1973 | 1974 | 1973 |
| Married | 52 | 61 | 60 | 35 | 67 |
| Not married | 56 | 67 | 63 | 46 | 66 |
| Full-time employed | 54 | 63 | 57 | 40 | 66 |
| Part-time employed | 56 | 64 | 51 | 42 | 70 |
| Not in employment | 55 | 68 | 69 | 40 | 64 |
| Partner full-time employed | 52 | 63 | 63 | 40 | 71 |
| Partner part-time employed | 47 | 68 | 50 | 43 | 48 |
| Partner not in employment | 56 | 64 | 65 | 21 | 65 |
| Children yes | 54 | 64 | 62 | 40 | 66 |
| Children no | 55 | 66 | 61 | 40 | 65 |

*Data source:* Web survey, Split 2; multiple coding possible for all categories but no generalization possible, other answers, and probe nonresponse, i.e. figures do not add up to 100%.

*Table A8* Answers to category-selection probe for respondents who opted for "individual-solution category" for closed question in the different countries, separately for male and female respondents (in percent)

| | Germany | | Great Britain | | United States | | Mexico | | Spain | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female |
| Time availability | 7 | 12 | 1 | 3 | 1 | 4 | 13 | 12 | 15 | 18 |
| Resource dependency | 17 | 24 | 16 | 17 | 9 | 7 | 8 | 13 | 13 | 14 |
| Individual preferences | 2 | 13 | 2 | 11 | 3 | 4 | 2 | 4 | 7 | 3 |
| Individual capabilities | 0 | 0 | 10 | 6 | 3 | 2 | 4 | 6 | 6 | 1 |
| Family/partners have to decide | | | | | | | | | | |
| - general | 5 | 13 | 3 | 6 | 6 | 6 | 2 | 1 | 0 | 2 |
| - no interference | 3 | 2 | 9 | 11 | 6 | 15 | 2 | 0 | 0 | 1 |
| - joint decision | 3 | 2 | 7 | 6 | 9 | 12 | 19 | 9 | 8 | 8 |
| Situation dependency | | | | | | | | | | |
| - financial necessities | 10 | 17 | 9 | 9 | 9 | 15 | 10 | 19 | 14 | 21 |
| - presence of one parent at home | 8 | 8 | 9 | 9 | 9 | 7 | 4 | 7 | 8 | 5 |
| Alternative possibilities | 12 | 14 | 1 | 8 | 7 | 7 | 6 | 13 | 15 | 20 |
| No generalization possible | 32 | 23 | 44 | 33 | 33 | 26 | 35 | 33 | 33 | 33 |
| Other answers | 13 | 0 | 6 | 7 | 17 | 14 | 13 | 12 | 7 | 5 |
| Probe nonresponse | 2 | 5 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| *N* | 60 | 84 | 68 | 96 | 70 | 98 | 48 | 69 | 86 | 87 |

*Data source:* Web survey, Split 2; multiple coding possible for all categories but no generalization possible, other answers, and probe nonresponse, i.e. figures do not add up to 100%

methods, data, analyses | Vol. 14(2), 2020, pp. 241-250 DOI: 10.12758/mda.2020.04

# The Effects of an Incentive Boost on Response Rates, Fieldwork Effort, and Costs across Two Waves of a Panel Study

*Katherine A. McGonagle*
*University of Michigan, Institute for Social Research*

**Abstract**

This paper describes the association between an incentive boost and data collection outcomes across two waves of a long-running panel study. In a recent wave, with the aim of achieving response rate goals, all remaining sample members were offered a substantial incentive increase in the final weeks of data collection, despite uncertainty about potential effects on fieldwork outcomes in the following wave. The analyses examine response rates and the average number of interviewer attempts to complete the interview in the waves during and after the incentive boost, and provide an estimate of the cost of the incentives and fieldwork in the waves during and following the boost. The findings provide suggestive evidence that the use of variable incentive strategies from one wave to the next in the context of an ongoing panel study may be an effective strategy to reduce nonresponse and may yield enduring positive effects on subsequent data collection outcomes.

*Keywords*: data collection, incentives, nonresponse, response rate, contact strategies, fieldwork effort, panel study

This paper examines the use of an increase in study incentives near the end of the field period in a recent wave of the Panel Study of Income Dynamics (PSID), a long-running household panel study of U.S. families, and the association with data collection outcomes, including respondent cooperation, fieldwork effort (as assessed by number of interviewer attempts to complete the interview), and fieldwork costs in the following wave.

The beneficial effects of providing incentives in exchange for participation in interviewer-administered surveys are well documented (e.g. see Laurie & Lynn, 2009; Singer & Ye, 2013). Substantial research based on longitudinal studies finds that incentives are associated with higher response rates (e.g., Fumagalli, Laurie, & Lynn, 2010; Hsu, Schmeiser, Haggerty et al., 2017; Martin, Abreu, & Winters, 2001; McGonagle & Freedman, 2017; McGonagle, Couper, & Schoeni, 2011; McGonagle, Schoeni, & Couper, 2013; Rodgers, 2002) and fewer attempts to complete an interview in the wave they are offered (e.g., Markesich & Kovac, 2003; McGonagle et al., 2013).

Despite numerous studies on the effects of incentives, the topic of differential incentive strategies in the context of ongoing panel studies has received little attention (see Singer & Ye, 2013). A handful of studies have found that incentives provided in a study's first wave have enduring effects on panel retention in subsequent waves (e.g., Goldenberg, McGrath, & Tan, 2009; James, 1997; Lengacher, Sullivan, Couper, et al., 1995; Mack, Huggins, Keathley, et al., 1998; McGrath, 2006; Pforr, Blohm, Blom, et al., 2015; Singer, Van Hoewyk, & Maher, 1998). While these findings indicate that the positive effects of incentives offered at study entry may persist across waves, it is unclear whether this applies to incentives offered later in a panel's history. In particular, the consequences of providing variable incentive amounts across sample members, or temporarily increasing incentive amounts within a particular wave – on data collection outcomes in future waves – are largely unknown.

During 2015, a differential incentive strategy was implemented in the PSID. As with panel studies across the world (De Leeuw, Hox, & Luiten, 2018), in recent waves PSID has experienced increased difficulty making contact with sample members and gaining their cooperation to complete the interview. In 2015, the study was faced with a substantially higher number of attempts by interviewers to make contact with sample members compared to prior waves, resulting in a high

*Acknowledgements*

This work was supported by the National Science Foundation [SES 1623864], the National Institute on Aging [R01 AG040213], and the National Institute of Child Health & Human Development [R01 HD069609].

*Direct correspondence to*

Katherine A. McGonagle, University of Michigan, Institute for Social Research
E-mail: kmcgon@umich.edu

proportion of outstanding sample at risk for nonresponse late in the field period. In the final weeks of data collection, a substantial incentive increase was offered to all remaining sample members. This strategy was undertaken to maintain the study's high response rate in the current wave, despite uncertainty about the impact on data collection outcomes in the following wave when the incentive was returned (i.e., reduced) to the baseline amount.

This paper examines the overall utility of the incentive boost across two waves of data collection in the PSID. The goal is to contribute to the "urgent need" identified by Laurie and Lynn (2009) to "extend the research knowledge base… to use survey budgets effectively and wisely when choosing respondent incentive strategies for longitudinal surveys." Using observational panel data, the following questions are considered: Is there evidence that a large incentive boost reduces nonresponse in the wave it is provided? What are the data collection outcomes in the wave following an incentive boost, when the incentive is returned to the baseline amount, including response rate and average number of interviewer attempts to complete the interview, and what percentage of respondents respond to the initial incentive, and what percentage respond only when the incentive is increased? Finally, the cost implications of the incentive boost are examined. What were the relative costs of the increased incentive in the current wave, and did these higher costs endure in the following wave? Limitations for the findings and next steps for research are described.

## Methods

This report draws on production data collected during the 2015 and 2017 waves of the Panel Study of Income Dynamics (PSID). The PSID is a longitudinal study of a nationally representative sample of U.S. families that began in 1968 and collects a variety of data on economic, health, and social behavior (see McGonagle, Schoeni, Sastry et al., 2012 for more information). Interviews have been conducted annually 1968-1997 and biennially since 1999 by professional interviewers employed by the Survey Research Operations group at the Survey Research Center within the Institute for Social Research at the University of Michigan. The study has achieved high wave-to-wave re-interview response rates, exceeding 93% in most waves. Data collection occurs in odd-numbered years between about March 1 and December 31 over the course of 44 weeks. December 31 is a firm end date for the collection of data each wave because the instrument questionnaire content focuses on specific time periods within the current calendar year.

Since 2003, the mode of data collection for approximately 97% of the sample has been computer-assisted telephone interview with in-person visits made to a small fraction of sample members. The study interviews one adult respondent in

each family, typically the individual who is most knowledgeable about the family finances (known as the "Reference Person"). Interviewers attempt to contact respondents primarily using telephone (comprising more than three-quarters of all contact attempts in 2015 and 2017), as well as by sending a small number of email and text messages. The average interview length was about 75 minutes in both 2015 and 2017. During 2015 and 2017, interviews were completed with 9,048 and 9,155 families with overall wave-to-wave re-interview response rates (i.e., response rates among those who had participated in the prior wave) of 93% and 94%, respectively.

*Use of incentives.* Since the inception of the study, post-paid monetary incentive payments have been offered to respondents in exchange for the completion of an interview. The incentive payment is typically provided by bank check to the family member who completes the interview. The general strategy in selecting the incentive amount is to offer an amount that roughly aligns with the interview length (i.e., roughly $1 USD for each minute of content) and to maintain a static amount for two waves that is modestly raised every third wave. These increases are intended to adjust for inflation and any increase in the length or general burden of the survey request. Sample members are provided with advance notice of the incentive amount being offered to complete the interview in an informational letter sent prior to the start of each wave of data collection. All subsequent messages sent to sample members requesting their participation reference the incentive. Historically, the incentive offer has remained unchanged throughout a wave of data collection, and all sample members have been offered the same incentive amount. In 2015, a baseline incentive of $70 USD was offered to 8,889 families who also participated in the prior wave (i.e., "re-interview cases").

During 2015, nearly 15% (1,322 cases) of the 8,889 re-interview cases had not completed their interview with approximately six weeks remaining in the production period. Reflecting the growing difficulty in recent waves of making contact with sample members in telephone studies, by comparison, with the same amount of time remaining in the 2013 wave, a much smaller fraction (6.6%) had not completed their interview. With the goal of achieving the target response rate for the 2015 wave, all remaining cases were offered a large incentive increase from $70 USD to $150 USD. The selection of the amount of the incentive increase was to make the survey request highly salient and reduce perceived barriers to participation by the study's end date. The incentive boost was communicated to respondents in various ways, including an announcement through a postcard sent via U.S. postal mail, through messages left by interviewers on telephones and cell phones, and through an email and text message. The $150 USD incentive remained in effect throughout the remaining weeks of the field period.

At the start of data collection the following wave (2017), the baseline incentive offer was restored. In this wave, the baseline incentive offer was $75 USD, an increase of $5 USD over the $70 USD baseline incentive offered at the start of 2015,

following the convention of modest increases in the baseline incentive every third wave. At the end of the 2017 field period with six weeks remaining, the incentive offer was again increased to $150 USD for all remaining sample members.

## Results

Table 1 presents response rates and field effort in the current and subsequent waves for respondents who were offered the incentive boost in the final six weeks of production during 2015. Field effort is defined as the average number of total attempts by the interviewer using telephone, email and text message required to complete the interview. The first column provides information on the fieldwork outcomes in the 2015 wave ("Current wave") for the 1,322 cases offered the 2015 incentive boost.

As shown in Table 1, the $150 USD incentive boost in 2015 had a positive impact on study participation with the majority of respondents (59.9%) completing the interview by the end of the field period, allowing response rate goals to be met. The second column provides information on fieldwork outcomes in the 2017 wave ("Next wave") for the subset of respondents who completed the 2015 interview after being offered the incentive boost. The key question is whether data collection outcomes for those now being offered $75 USD to complete their interview, half as much, were negatively affected. The results show that there is no evidence that respondents were reluctant to participate given the reduced incentive amount. The vast majority of respondents – nearly 89% – who received the $150 USD incentive

*Table 1* Fieldwork outcomes over two waves for re-interview respondents offered an incentive boost

| | Current wave | Next wave |
|---|---|---|
| | 2015 (n=1,322) | 2017 (n=780) |
| Response rate[1] | 59.9% | 88.6% |
| Number of attempts among respondents (mean) | 82.8 | 33.7 |
| Incentive amount required for response | | |
| $150 (boost) | 100.0% | |
| $75 (baseline offer) | | 73.3% |
| $150 (end of study offer) | | 26.7% |
| Total | 100.0% | 100.0% |

[1] Of the 791 repondents who completed the 2015 interview following the $150 USD incentive boost, 11 were ineligible for the study 2017

boost in 2015 continued to participate during the 2017 wave. Moreover, field effort in the 2017 wave actually decreased substantially for those receiving the incentive boost compared to the 2015 wave, dropping from an average number of 82.8 interviewer attempts to complete the interview in 2015 to an average of 33.7 interviewer attempts in 2017.

A second key question is what proportion of respondents who received the incentive boost in 2015 completed the interview in 2017 for the baseline incentive of $75 USD, and what proportion delayed participation until being offered $150 USD. As shown in the table, the vast majority of these respondents – 73.3% – completed their 2017 interview for the baseline incentive offer of $75 USD. Another 26.7% of those who required $150 USD to respond in 2015 responded in 2017 only after again being offered $150 USD near the close of the field period.

Among those completing their interview for the $75 USD baseline incentive, the average number of interviewer attempts was only about 16.0, compared to about 65.0 interviewer attempts on average for those cases who again delayed their participation for the $150 USD at the end of the 2017 field period (not shown in table).

The final question considers the cost-implications of the incentive boost. A concern for survey organizations is that respondents who receive an incentive increase in one wave may resist completing the interview if offered a lower amount in a future wave, leading such increases to be permanent. Did the 2015 incentive boost lead to enduring costs in the following wave? A basic estimate of the fieldwork effort and incentive costs in each wave for the 780 respondents who participated in both waves was generated. A cost-per-interviewer-attempt estimate of $5.50 USD was derived based on the average hourly wage of an interviewer ($22 USD) and the assumptions that interviewers could make four attempts per hour and that each attempt type (telephone, email and text message) required the same amount of time ($22 USD/4 attempts = $5.50 USD). As shown in Table 2, using the average number of interviewer attempts across the 780 cases (i.e., average attempts of 82.8 in 2015 and 33.7 in 2017), fieldwork costs for these respondents are estimated at $355,212 USD in 2015 and at $144,573 USD in 2017. Incentive costs in 2015 were $117,000 USD (i.e., all 780 respondents required $150 USD). In 2017 incentive costs for these 780 respondents dropped by more than one-third to $74,120 USD (i.e., 73.3% responding during the baseline offer of $75 USD and 26.7% responding for the increased offer of $150 USD). Summing costs attributable to fieldwork effort and incentive payments yields total costs of $472,212, or $605 per case in 2015, and $218,693 or $280 per case in 2017, a decline of more than 50% in total costs. In sum, both incentive costs and fieldwork costs decreased substantially for cases receiving the increased incentive in the subsequent wave.

*Table 2* Cost estimates of fieldwork effort by wave

| | Current wave | Next wave |
|---|---|---|
| Cost parameters | 2015 | 2017 |
| Number of cases responding in both waves | 780 | |
| Average cost per interviewer attempt | $5.50 | |
| Total interviewer attempts (mean) | 82.8 | 33.7 |
| Average cost of interviewer attempts | $355,212 | $144,573 |
| Average cost of incentive payments | $117,000 | $74,120 |
| Total cost | $472,212 | $218,693 |
| Cost per case | $605 | $280 |

# Discussion

The goals of the current study were to examine the effects of an increased incentive on cooperation late in the field period of a long-running panel study, and trace its effects to response rates and fieldwork outcomes in the following wave. An important limitation to note at the outset is the lack of a randomly selected control group in the assignment of the incentive boost. Since all late-responding sample members were offered an increased incentive, it is not possible to compare outcomes with those who were not offered a higher incentive amount. A second limitation is that the results of the current study are drawn from the experience of a specific ongoing panel study comprising U.S. adults whose families have participated across many decades, making the generalizability of the results to other study designs uncertain.

Despite these limitations, several key findings have emerged from this descriptive analysis. First, the incentive boost was successful in achieving the main operational goal of meeting response rate targets in the wave it was implemented, inducing cooperation from a high percentage of respondents late in the field period. Second, there is no evidence that the increased incentive negatively affected data collection outcomes among respondents offered a lower initial incentive in the subsequent wave, with nearly 89% completing an interview. Moreover, those receiving the incentive boost required substantially less field effort in the following wave to complete their interview than was needed to finalize their interview in the wave they received the boost. Third, contrary to the concern that the costs of the incentive boost would endure in the subsequent wave, costs substantially declined, with

the majority of respondents completing the 2017 interview for the baseline offer with about one-third fewer contact attempts than needed in the prior wave.

In providing suggestive evidence that the positive effects of monetary incentives may persist over time, this descriptive analysis is consistent with the handful of studies on this topic in the literature (Jäckle & Lynn, 2008; Mack et al., 1998; Scherpenzeel, Zimmermann, & Budowski, 2002). In the current study, the concern that those who were offered a substantially higher incentive at a point in time would then delay their participation until the same amount was offered was not realized for the majority of respondents.

In the context of a long-running panel study, the offer of a substantial incentive increase may induce survey participation by highlighting to respondents the legitimacy of the study and the value of their participation. Moreover, interviewers likely gain confidence from the raised incentive when making contact with "difficult" respondents who have evaded many prior attempts. Such mechanisms have been suggested to also underlie the beneficial impact of respondent materials, such as letters sent by survey organizations in advance of data collection (De Leeuw, Callegaro, & Hox, 2007). These positive effects may carry-over to subsequent requests for survey participation, potentially by building rapport and good-will, as well as through the elicitation of principles of social exchange and reciprocation (see e.g., Dillman, Smyth, & Christian, 2009).

A note on the choice of the amount of the incentive increase is in order. In the selection of initial monetary incentive amounts and subsequent magnitudes of increases that may occur during fieldwork, survey practitioners have little research evidence on which to draw. This can be traced to the challenges of mounting experiments during active data collection which may have uncertain effects on study goals, as well as the highly contextualized nature of study designs where multiple factors must be considered in the selection of incentive amounts, including respondent characteristics, interview length and burden, and budgetary constraints. Our goal was to implement a highly salient incentive increase in order to reduce respondent barriers to participation and achieve a particular response rate goal by the firm end date of the study. Designing and implementing experimental studies on this topic to better understand the relative effectiveness of different orders of magnitudes of incentive increases would be of high value to the field.

In summary, the findings of this study are consistent with prior research documenting the positive effects of incentives on data collection outcomes. The results additionally provide suggestive evidence that using variable incentive strategies over waves of fieldwork in the context of a large national panel study may be an effective strategy to maximize response rates and yield enduring positive effects on subsequent participation and field effort. An important consideration for ongoing panel studies in future research is how individual characteristics of sample members may affect responsivity to differential incentives and influence sample

bias over subsequent waves. Future research should replicate these findings using experimental methods to better understand the mechanisms through which these outcomes occur.

## References

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E., & Lensvelt-Mulders, G. (2007). The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis. *Public Opinion Quarterly 71*(3), 413-443.

De Leeuw, E., Hox, J., & Luiten, A. (2018). International Nonresponse Trends across Countries & Years: An Analysis of 36 Years of Labour Force Data. Survey Insights: Methods from the Field. Retrieved from https://surveyinsights.org/?p=10452.

Dillman, D.A., Smyth, J.D., & Christian, L.M. (2009). *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method. 3rd ed.* New York, NY: Wiley.

Fumagalli, L., Laurie, H., & Lynn, P. (2010). Experiments with Methods to Reduce Attrition in Longitudinal Surveys. Institute for Social and Economic Research Working Paper 2010-04. University of Essex. https://www.iser.essex.ac.uk/files/iser_working_papers/2010-04.pdf

Goldenberg, K. L., McGrath, D., & Tan, L. (2009). The Effects of Incentives on the Consumer Expenditure Interview Survey. In JSM Proceedings, 5985-99. Alexandria, VA: American Statistical Association. https://www.bls.gov/osmr/research-papers/2009/st090100.htm

Hsu, J. W., Schmeiser, M. D., Haggerty, C., & Nelson, S. (2017). The Effect of Large Monetary Incentives on Survey Completion: Evidence from a Randomized Experiment with the Survey of Consumer Finances. *Public Opinion Quarterly 81*(3), 736-747. https://academic.oup.com/poq/article/81/3/736/3798583

Jäckle, A. & Lynn, P. (2008). Respondent Incentives in a Multi-mode Panel Survey: Cumulative Effects on Nonresponse and Bias. S*urvey Methodology 34*(1), 105–17.

James, T. (1997). "Results of the Wave 1 Incentive Experiment in the 1996 Survey of Income and Program Participation." Proceedings of the Survey Research Section of the American Statistical Association.

Laurie, H., & Lynn, P. (2009). The Use of Respondent Incentives on Longitudinal Surveys. In Lynn, P., *Methodology of Longitudinal Surveys*. London: John Wiley.

Lengacher, J. E., Sullivan, C. M., Couper, M. P., & Groves, R. M. (1995). "Once Reluctant, Always Reluctant? Effects of Differential Incentives on Later Survey Participation in a Longitudinal Study." Proceedings of the Section on Survey Methodology, American Statistical Association, 1029-34. http://www.asasrms.org/Proceedings/papers/1995_179.pdf

Mack, S., Huggins, V., Keathley, D., & Sundukchi, M. (1998). "Do Monetary Incentives Improve Response Rates in the Survey of Income and Program Participation?" Proceedings of the Section on Survey Methodology, American Statistical Association, 529-34. http://www.asasrms.org/Proceedings/papers/1998_089.pdf

Markesich, J., & Kovac, M. D. (2003). The Effects of Differential Incentives on Completion Rates: A Telephone Experiment with Low-Income Respondents. Paper presented at the annual meeting of the American Association of Public Opinion Research, May 16, 2003, Nashville, TN.

Martin E., Abreu, D., & Winters, F. (2001). Money and Motive: Effects of Incentives on Panel Attrition in the Survey of Income and Program Participation. *Journal of Official Statistics, 17*(2), 267-284. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/money-and-motive-effects-of-incentives-on-panel-attrition-in-the-survey-of-income-and-program-participation.pdf

McGonagle, K. A., Couper, M. P., & Schoeni, R. F. (2011). Keeping Track of Panel Members: An Experimental Test of a Between-Wave Contact Strategy. *Journal of Official Statistics, 27*(2), 319-338.

McGonagle, K. A., Schoeni, R. F., Sastry, N., & Freedman, V. A. (2012). The Panel Study of Income Dynamics: Overview, Recent Innovations, and Potential for Life Course Research. *Longitudinal and Life Course Studies, 3*(2), 268-284.

McGonagle, K. A., Schoeni, R. F., & Couper, M. P. (2013). The Effects of a Between-Wave Incentive Experiment on Contact Update and Production Outcomes. *Journal of Official Statistics, 29*(2), 1-17.

McGonagle, K. A. & Freedman, V. A. (2017). The Effects of a Delayed Incentive on Response Rates, Response Mode, Data Quality, and Sample Bias in a Nationally Representative Mixed Mode Study. *Field Methods, 29*(3), 221-237.

McGrath, D. E. (2006). An Incentives Experiment in the U.S. Consumer Expenditure Quarterly Survey. In JSM Proceedings, 3411-18. Alexandria, VA: American Statistical Association. https://www.bls.gov/osmr/pdf/st060030.pdf

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., et al. (2015). Are Incentive Effects on Response Rates and Nonresponse Bias in Large-scale, Face-to-face Surveys Generalizable to Germany? Evidence from Ten Experiments. *Public Opinion Quarterly, 79* (3), 740–768. https://academic.oup.com/poq/article/79/3/740/1916249

Rodgers, W. (2002). Size of Incentive Effects in a Longitudinal Study. Proceedings of the Survey Research Methods Section of the American Statistical Association (pp. 2930-2935). Washington, DC: American Statistical Association. https://pdfs.semanticscholar.org/b8ec/5d6ef31b383739824edf4715ad50d413110a.pdf

Scherpenzeel, A., Zimmermann, E., Budowski, M., Tillmann, R., Wernli, B., Gabadinho, A. (2002). Working Paper, No. 5-02. Neuchatel: Swiss Household Panel. [Accessed May 31, 2012]. Experimental Pre-Test of the Biographical Questionnaire. Available at: http://aresoas.unil.ch/workingpapers/WP5_02.pdf. [Google Scholar]

Singer, E., Van Hoewyk, J., & Maher, M.P. (1998). Does the Payment of Incentives Create Expectation Effects? *Public Opinion Quarterly 62*, 152–64. https://www.ssoar.info/ssoar/bitstream/handle/document/49721/ssoar-1998-singer_et_al-Does_the_payment_of_incentives.pdf?sequence=1

Singer, E. & Ye, C. (2013). The Use and Effects of Incentives in Surveys. *The Annals of the American Academy of Political and Social Science, 645*(1), 112-141. https://journals.sagepub.com/doi/full/10.1177/0002716212458082

# Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

**Entire Book:**

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

**Journal Article (with DOI):**

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

**Journal Article (without DOI):**

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

**Chapter in an Edited Book:**

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

**Internet Source (without DOI):**

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)