The Relationship Between Response Probabilities and Data Quality in Grid Questions

Gummer, Tobias; Bach, Ruben L.; Daikeler, Jessica; Eckman, Stephanie

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Survey Research Methods (2021)
Vol. 15, No. 1, pp. 65-77
doi:10.18148/srm/2021.v15i1.7727
European Survey Research Association

The relationship between response probabilities and data quality in grid questions

Tobias Gummer
GESIS – Leibniz-Institute for the Social Sciences
Mannheim, Germany

Ruben Bach
University of Mannheim
School of Social Sciences
Mannheim, Germany

Jessica Daikeler
GESIS – Leibniz-Institute for the Social Sciences
Mannheim, Germany

Stephanie Eckman
RTI International
Washington DC, USA

Response probabilities are used in adaptive and responsive survey designs to guide data collection efforts, often with the goal of diversifying the sample composition. However, if response probabilities are also correlated with measurement error, this approach could introduce bias into survey data. This study analyzes the relationship between response probabilities and data quality in grid questions. Drawing on data from the probability-based GESIS panel, we found low propensity cases to more frequently produce item nonresponse and nondifferentiated answers than high propensity cases. However, this effect was observed only among long-time respondents, not among those who joined more recently. We caution that using adaptive or responsive techniques may increase measurement error while reducing the risk of nonresponse bias.

Keywords: response propensity; measurement error; data quality; panel survey; adaptive survey design

1 Introduction

Response propensities are estimates of a sampled individual's probability to respond to a survey. They are often used in adaptive and responsive survey designs to guide data collection efforts (Schouten, Peytchev, & Wagner, 2017; Tourangeau, Michael Brick, Lohr, & Li, 2017; Wagner, 2008). For example, a common approach uses response propensities to tailor interventions and balance the sample (Peytchev, Riley, Rosen, Murphy, & Lindblad, 2010; Rosen et al., 2014) to reduce the risk of nonresponse bias (Schouten, Cobben, Lundquist, & Wagner, 2016). The U.S. National Survey of Family Growth, for instance, instructed interviewers to give priority to cases with low response propensities as those are hardest to recruit for a survey interview (Wagner et al., 2012). The German Longitudinal Election Study experimented with assigning low response propensity cases to the most experienced interviewers in order to make sure that these reluctant cases become actual respondents (Gummer & Blumenstiel, 2018). However, if response probabilities are associated with other sources of error (e.g., Groves, 2006; Tourangeau, 2019), these interventions may lead to unintended consequences. In particular, if probabilities are correlated with measurement error, prioritizing cases by propensity may impact not only nonresponse bias and outcome rates but also measurement error and total survey error (Bach, Eckman, & Daikeler, 2020; Kreuter, Müller, & Trappmann, 2010). For instance, Peytchev, Peytcheva, and Groves (2010) found low propensity cases to be more likely to misreport prior experiences. Similarly, Fricker and Tourangeau (2010) found low propensity respondents more likely to report rounded, less precise values to continuous variables, and more prone to item nonresponse.

This study tests the relationship between response probabilities and measurement error in a probability-based panel survey in Germany. Since response probabilities are latent constructs with unknown true values, they have to be estimated, for example, via logistic regression modeling or nonparametric prediction techniques based on classification tree algorithms (Kern, Klausch, & Kreuter, 2019). Panel surveys are an ideal setting for such estimation, because they often come with a rich set of historical data about cases from previous waves which can be used to obtain accurate estimates of the response probabilities.

Bethlehem, Cobben, and Schouten (2011, p. 44) show

Contact information: Tobias Gummer, GESIS – Leibniz-Institute for the Social Sciences, PO Box 12 21 55, 68072 Mannheim, Germany (E-mail: tobias.gummer@gesis.org)

that the nonresponse bias of the sample mean (i.e., based on those respondents who participate in a survey) of variable Y ($\bar{y}_R$) is approximately equal to

$$B(\bar{y}_R) = \frac{R_{\theta Y} S_\theta S_Y}{\bar{\theta}} \quad (1)$$

where $R_{\theta Y}$ is the correlation between the response probability θ and Y, S_θ is the standard deviation of θ, S_Y is the standard deviation of Y, and $\bar{\theta}$ is the mean response probability. The formula shows that a nonresponse bias exists, if response probabilities vary in the target population and are correlated with the variable of interest. In this regard, Groves (2006, p. 651–652) describes the "Nonresponse-Measurement Error model" in which a measurement error ϵ is caused by θ. If this relationship exists, and $R_{\theta Y} \neq 0$ and $S_\theta \neq 0$, $\bar{y}_R$ will further be affected by measurement error.

We focus our assessment of measurement error on grid questions because they are burdensome and can result in increased levels of measurement error. For example, grid questions lead to less differentiated responses, more speeding, and less substantive answers compared to item-by-item designs (Liu & Cernat, 2018; Roßmann, Gummer, & Silber, 2018; Tourangeau, Couper, & Conrad, 2004; Tourangeau, Maitland, et al., 2017). Despite these issues, grid questions are still frequently used, because they reduce interview duration and require less space. Their susceptibility to quality issues makes grid questions an important case study of the relationship between response probabilities and measurement error.

In the next section we present our data and methods. We then discuss our results before concluding with practical implications of our findings and future research opportunities.

2 Data and Methods

To investigate the relationship between response probabilities and measurement error, we draw on data from the GESIS panel (GESIS Panel Team, 2018), a probability-based mixed-mode general population panel in Germany (Bosnjak et al., 2018).[1] In 2013, the panel was recruited via face-to-face interviews with a sample drawn from population registers. The recruitment survey achieved a Response Rate 5 (AAPOR, 2016) of 38.6%, and 4,888 panelists joined the panel in 2014. We refer to this sample as the initial sample. To account for panel attrition and aging, a refreshment sample was drawn as part of the German General Social Survey (ALLBUS) in 2016. Similar to the initial sample, the ALLBUS drew a sample from population registers aiming to cover the German population. Thus, the refreshment sample covered the full target population of the GESIS panel. The ALLBUS achieved a Response Rate 5 of 34.8% and 1,710 respondents joined the GESIS panel. We refer to this sample as the refreshment sample.

Upon recruitment, the panelists were asked about their preferred mode of participating in re-interviews: web or mail. Every two months, all active panelists are interviewed using their preferred mode. Changing the mode of participation later is possible by contacting the fieldwork institute. However, few respondents have made use of this possibility. The survey length is limited to approximately 20 minutes per wave. Questions are presented similarly in the two modes to minimize mode effects. Respondents who do not participate in three waves in a row are dropped from the panel and no longer receive invitations to participate.

For the present study, we use wave "ed," the fourth wave in 2017, because this wave featured an attention check item in one of its grids, which we use to assess respondent attentiveness. Attention checks are instruments designed to measure whether respondents answer questions without thoroughly processing them (Gummer, Roßmann, & Silber, 2021; Meade & Craig, 2012). Frequently, these checks involve a task that respondents have to complete to indicate their attentiveness, for example "click strongly agree" or "instead of answering click the blue box in the right corner". Overall, nine grids were included in the wave's questionnaire, one of which contained an attention check. All scales were presented in a horizontal orientation. An example is provided in the Appendix, Figure A1.

All 4,777 active panelists were invited to participate in wave "ed" (initial sample = 3,287; refreshment sample = 1,490) of which 4,257 completed the questionnaire (initial sample = 2,976; refreshment sample = 1,281). Prior to wave "ed," panelists of the initial sample had been invited to participate in 22 waves, whereas panelists of the refreshment sample had only been invited to between 4 and 7 waves, depending on when they were recruited in 2016. Of the respondents who completed the survey, 33% participated via mailed questionnaire and 67% via the web (initial sample = 68% web; refreshment sample = 65% web). Among the web respondents, 15% completed the survey using mobile devices.

2.1 Data quality measures

To investigate measurement errors in wave "ed," we use six data quality measures for each of the nine grid questions. We focus on measures that indicate whether respondents may have skipped steps in the cognitive answering process (Tourangeau, Rips, & Rasinski, 2000) to reduce their response burden (Krosnick, 1991, 1999).

Left-aligned responses. The proportion of rows in a grid where the left response option was used. Selecting the first available answer on a scale can be a strategy to re-

[1] Data used in our study are available at the GESIS Data Archive: study number ZA5665, Version 24.0.0 (GESIS Panel Team, 2018).

duce response burden (Krosnick, 1999). This variable takes values between 0 and 1, inclusive.

Extreme responses. The proportion of rows in a grid where a scale end-point was selected. To reduce the complexity of the response process, respondents could align their answers to scale endpoints (Paulhus, 1991). This variable takes values between 0 and 1, inclusive.

Item Nonresponse. The proportion of rows in a grid where no answer was given. Respondents were able to skip any row to ease their response process (Krosnick, 1991). This variable takes values between 0 and 1, inclusive.

Straightlining. Indicator (0,1) that all rows in a grid have the same answer. Satisficing response behavior (Krosnick, 1991, 1999) can lead respondents to base their answers to a grid on the first response option they deem satisfactory, if they then align their subsequent responses to the grid to their first response, straightlining occurs (e.g., Roßmann et al., 2018).

Probability of differentiation (ρ). The use of diverse response options. As an additional measure for nondifferentiation in responses to a grid, we calculate the probability of differentiation ρ, as suggested by Krosnick and Alwin (1988). ρ takes values between 0 and 1 and is defined as: $\rho = 1 - \sum_{i=1}^{n} P_i^2$, where n is the number of response options and P_i is the proportion of rows using response option i. ρ is missing for all grids where any row has a missing value. Higher values indicate the use of a more diverse set of response options and 0 indicates straightlining.

Coefficient of variation (CV). The variation in chosen response options. To complement ρ, we compute an additional measure of nondifferentiation: $CV = \frac{S}{\overline{x}}$ where $\overline{x}$ denotes the mean of all grid responses and S is the standard deviation of these answers. CV is set to missing for all grids where any item has a missing value. Similar to ρ, 0 indicates straightlining and higher values indicate more differentiation and potentially higher data quality (McCarty & Shrum, 2000).

Failing an attention check. The eighth grid in the wave "ed" questionnaire contained an instructed response item attention check (Gummer et al., 2021; Meade & Craig, 2012). Respondents were instructed to select the response option "rather disagree" for the sixth item. The grid was part of an experiment; two-thirds of the respondents were randomly assigned to receive the attention check item. This measure is at the case level and is 0 or 1 where 1 means that the respondent gave the wrong answer.

The first six measures are at the case-grid level; each respondent has a score for each measure for each of the nine grids. The seventh measure, the attention check question, is at the case level, though only two-thirds of the respondents received the attention check question. The final data set contains 38,313 data points clustered in 4,257 respondents.

To ease interpretation and comparison, we rescale all measures to take values between 0 and 1, where 0 indicates low data quality in a grid (e.g., providing item nonresponses, nondifferentation) and 1 indicates high data quality. Table 1 gives summary statistics for the seven measures after rescaling. Correlations between the seven data quality measures are shown in Figure 1. Some correlations are quite high, especially between the three measures of nondifferentiated responding (straightlining, CV, ρ). Others, however, are less than .1, indicating that the measures capture different aspects of measurement error.

2.2 Estimation of response probabilities

Response propensities are estimates of the probability that a case will respond to a survey (Bethlehem et al., 2011, Chapter 11). Response propensities vary between zero and one, with higher values indicating higher probabilities to respond to a survey. We estimate them with a model where the dependent variable is each invited panel member's observed response status in wave "ed". The independent variables are the characteristics of a person (such as their age, gender, and education) and information about the survey process in wave "ed" as well as data from previous waves (e.g., how many times a person was contacted and in which mode). A list of all variables used for the estimation of the response propensity model is given in Appendix Table B1.

Logistic regression is often used to estimate response propensity models, but prediction algorithms from the machine learning literature have received increased attention in the survey literature in recent years for several reasons (e.g., Bach et al., 2020; Buskirk & Kolenikov, 2015; Kern et al., 2019). Researchers using logistic regression models need to specify the functional form of their model in advance. That is, they need to choose which variables should be included in the model and in which functional form (e.g., in linear or quadratic form, with or without interactions). While such approaches are useful for hypothesis or theory testing, specifying a model in advance is less useful when the goal is to model the response process as closely as possible. With tree-based machine learning algorithms such as the one used in this paper, the response process modelling is completely driven by the associations found in the data, which often results in more accurate estimates of the response probabilities (see, e.g. Kern et al., 2019). That is, using a tree-based machine learning model, we do not need to specify the functional form of the response propensity model in advance. In addition, we do not need to decide which of the variables

Table 1
Descriptive statistics for Data Quality Measures

Data quality measure	Mean	Min	Max	N(grids)	N(obs)
Left-aligned responses	0.917	0	1	9	38,313
Extreme responding	0.814	0	1	9	38,313
Item Nonresponse	0.992	0	1	9	38,313
Straightlining	0.968	0	1	9	38,313
ρ	0.577	0	1	9	36,629
Coefficient of variation	0.340	0	1	9	36,629
Failing attention check	0.849	0	1	1	2,320

Note. All measures rescaled: 0 is low quality and 1 is high quality.

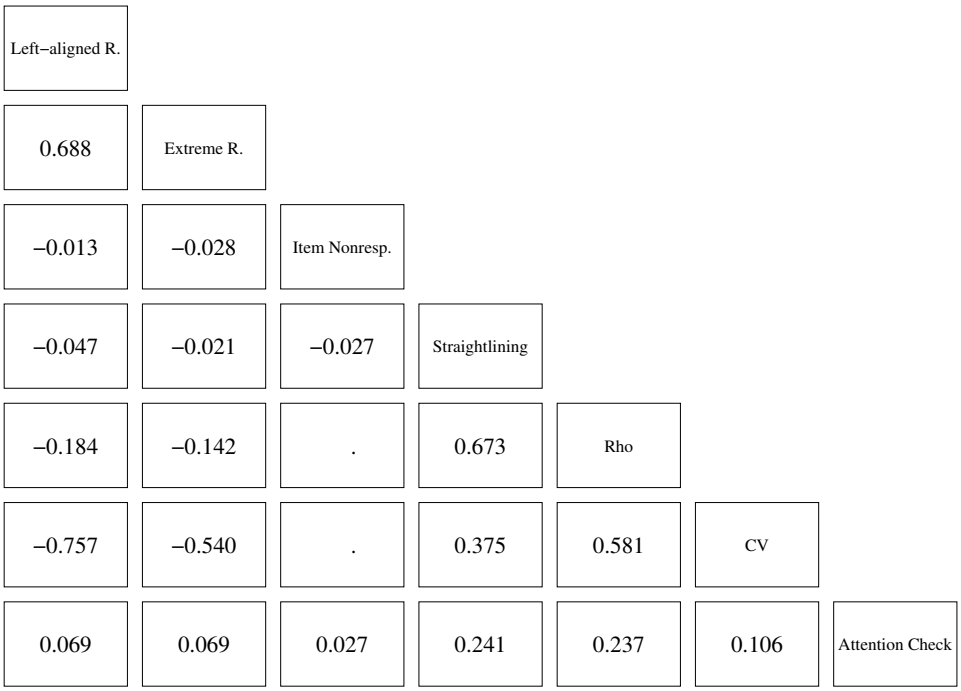

Figure 1. Correlation matrix of seven data quality measures

should be included in the model. Instead, we feed the algorithm all information available for both respondents and nonrespondents and let the algorithm decide which variables should enter the model and with what functional form. This decision is based only on the data, the explanatory power of the independent variables and the resulting performance of the model. Since we are interested in obtaining accurate estimates of the response probabilities and not in the model itself, this data-driven approach is preferable to the theory-driven approach underlying traditional regression modeling. Moreover, several studies have shown that such tree-based machine learning algorithms often outperform standard regression models in terms of bias in the resulting response propensities (see, e.g., Buskirk & Kolenikov, 2015; Kern et al., 2019).

In this study, we use the gradient boosting machine algorithm as implemented in the "gbm" package (version 2.1.3) in R (Friedman, 2001; R Core Team, 2020; Ridgeway, 2017). Briefly speaking, this algorithm is based on a combination of several classification trees.[2] Each classification tree works by splitting the predictor space (i.e., the set of all possible values of the predictors) into non-overlapping rectangular regions such that the the resulting regions are as homogeneous as possible with respect to the outcome variable (i.e., the response status of an individual). In other words, the data are split into regions such that the share of respondents in a resulting region is either very high or very low.

The tree-growing-process starts by determining the pre-

[2]Our description of the method draws heavily on Kern et al. (2019) and Bach et al. (2020).

dictor and its cut-point such that the resulting split creates two sub-regions where the homogeneity of the outcome is maximal. Through a recursive approach, the tree is then grown by considering each resulting sub-region for a new split using the process described above. Eventually, this process will result in sub-regions with only one observation. Such a tree, however, would perform poorly when applied to new data. Therefore, stopping criteria such as a minimum number of observations per sub-region may be applied.

Because a single tree often performs poorly (Kern et al., 2019), boosting uses a combination of several trees to form the final prediction model. Each tree is built from the results of the previous tree: new trees are fit to the difference between the observed outcome and the predicted probability. That is, a new tree tries to explain what the previous tree(s) could not. In this way, boosting algorithms aim to find a combination of trees such that each new tree adds an improvement to the previous tree and thereby improves the overall algorithm. In the end, this process results in a powerful combination of many trees. For further technical details, see, e.g. Friedman (2001), Kern et al. (2019), Ridgeway (2017).

Following standard practice, we randomly split the 4,777 invited cases into training (75%) and test (25%) sets (Hastie, Tibshirani, & Friedman, 2009, Chapter 7). We train and tune the boosting algorithm on the training sample using five-fold cross-validation to guard against overfitting. Final model performance is then evaluated on the test set to get a realistic estimate of the test error. The final response propensity model achieves an accuracy of 0.79. That is, we correctly classify 79 percent of all participants in the test set as either a respondent or a nonrespondent (using Youden's J statistic to determine the optimal probability cutoff (Youden, 1950). Our model seems to do equally well at identifying both respondents and nonrespondents: sensitivity is 0.80 and specificity is 0.77. Regarding the area under the (receiver operating characteristic) curve (AUC), our model achieves *excellent* discrimination between the two classes (AUC=0.84), according to the rules of thumb proposed by Hosmer and Lemeshow (2000). Moreover, the AUC tells us that the chance that a randomly chosen respondent has a higher response propensity score than a randomly chosen nonrespondent is 0.84. Thus, our model seems to predict response in wave "ed" of the panel survey very well, providing us with accurate estimates of each participant's response probability.

We then predict the response propensity for all responding cases (n = 4,257). Only responding cases are needed for analysis, because the data quality measures developed above are available only for those panelists who responded in wave "ed". The mean response propensity among respondents from the initial sample is 0.93 (Std. Dev. = 0.08, Min = 0.18, Max = 0.97). Response propensities among respondents in the refreshment sample are similar: mean of 0.92 (Std. Dev. = 0.08, Min = 0.18, Max = 0.97).

2.3 Analyses

To understand the relationship between response probabilities and data quality, we fit seven mixed models. In each model, the dependent variable is one of the measures of data quality. The independent variables are the predicted response propensity, an indicator for the nine grids, and an indicator for the refreshment sample. We include indicators for the grids (dummy variables) to control for differences between the grids regarding content and layout that may impact response behavior. In addition, we include an interaction effect between the sample type and the response propensities. We suspect that the relationship between response propensity and data quality might differ for respondents of the refreshment sample who have been invited to at most 7 waves and the initial sample who participated in up to 22 waves. The model also includes random intercepts at the respondent level to allow for the fact that data quality varies between respondents for reasons not captured in our independent variables.

All analyses are done using Stata 15.1. Replication code for the analyses and the prediction of response propensities is available in the online appendix.

3 Results

The mixed models reveal whether response propensities correlate with data quality measures in the initial and refreshment samples. Table 2 shows the estimated slope coefficients from the models separately for the two sample types: initial and refreshment. The third column tests for differences between the samples' slopes. Full results from the regression models are provided in the Appendix, Table B2. The model for the attention check question has many fewer cases because each respondent saw that question only once and one-third of the respondents did not see it at all.

For the initial sample, we find a significant relationship between data quality and response propensities for four measures: item nonresponse, straightlining, ρ, and the CV. *Higher* response propensities are associated with *better* data quality. High propensity respondents less frequently refused to answer when completing the grids of the questionnaire compared to low propensity respondents. Similarly, panelists with higher response propensities provide more differentiated answers compared to low propensity respondents. They use more of the available response options and their answers show more variation. Our data show no significant effects of response propensity on left-aligned responses, extreme responses, and failing the attention check. Interestingly, we find no significant associations between response propensities and data quality for the refreshment sample (see Table 2).

We then test whether the slopes differ significantly between the initial and refreshment samples. That is, we analyze whether the relationship between response propensities

Table 2
Regression estimates on relationship between response propensities and data quality measures by sample

	Sample		Difference between samples		
Data quality measure	Initial Coeff[a](SE)	Refreshment Coeff[a](SE)		N(grids)	N(respondents)
Left-aligned responses	−0.026 (0.016)	0.017 (0.020)		9	38313
Extreme responding	−0.033 (0.032)	0.057 (0.039)		9	38313
Item Nonresponse	0.016*** (0.004)	−0.005 (0.005)	**	9	38313
Straightlining	1.420* (0.686)	−0.014 (0.950)		9	38313
ρ	0.045* (0.018)	−0.021 (0.022)	*	9	36629[b]
Coefficient of variation	0.051** (0.018)	−0.016 (0.021)	*	9	36629[b]
Failing attention check	−0.259 (0.862)	0.552 (0.939)		1	2320

[a] Estimated relationship between response propensities and data quality measure
[b] N differs because of omission of missing values (see Data section)
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

and data quality measures vary between the samples. The slopes differ in the item nonresponse, ρ, and CV models. Figure 2 visualizes the slopes for these three data quality measures. The plots show that these differences in data quality between samples are most pronounced for lower propensities, whereas high propensity cases appear to provide responses of relatively similar quality.

4 Discussion and Conclusion

With the present study, we investigate the relationship between response probabilities and measurement error in grid questions in a panel survey. Drawing on data from the probability-based GESIS panel, we find response propensities to be associated with several measures of data quality. However, this relationship holds only for respondents who are part of the initial sample. We find no such effects for respondents who are part of the newly recruited refreshment sample, who have participated in fewer waves of the panel.

We can only speculate about the reasons for the differences in effects between the samples. In an systematic review, Olson (2013) summarizes different hypotheses on relationship between nonresponse (i.e., response probabilities) and data quality. She argues that the relationship is likely not explained by one single hypotheses but the interplay of different effects. In our view, this could be a possible explanation for our findings. On the one hand, the refreshment sample consists of newly recruited panelists who recently agreed to participate in re-interviews. These respondents can be reasoned to properly answer questions as they remember their participation decision well and consequently try to behave like a good respondent. Olson (2013) labels this explanation the "self-perception hypothesis" with reference to self-perception theory (Bem, 1967). We assume this effect to counteract other effects (e.g., motivation) that would result in a negative relationship between response probabilities and data quality. On the other hand, respondents in the initial sample participated for several waves and effects of decisions made during the panel recruitment should be less pronounced. Consequently, beneficial effects such as the self-perception hypothesis should be smaller and not able to counteract diminishing effects of factors that influence nonresponse and data quality.

We do not find effects for all data quality measures in our analyses. This finding highlights the importance of acknowledging that data quality encompasses different aspects that are the result of different response behaviors and motivations—some of which are related to response probabilities.

Our findings have practical implications for surveys that feature adaptive and responsive designs, particularly panel surveys. First, efforts to balance the sample, for instance by

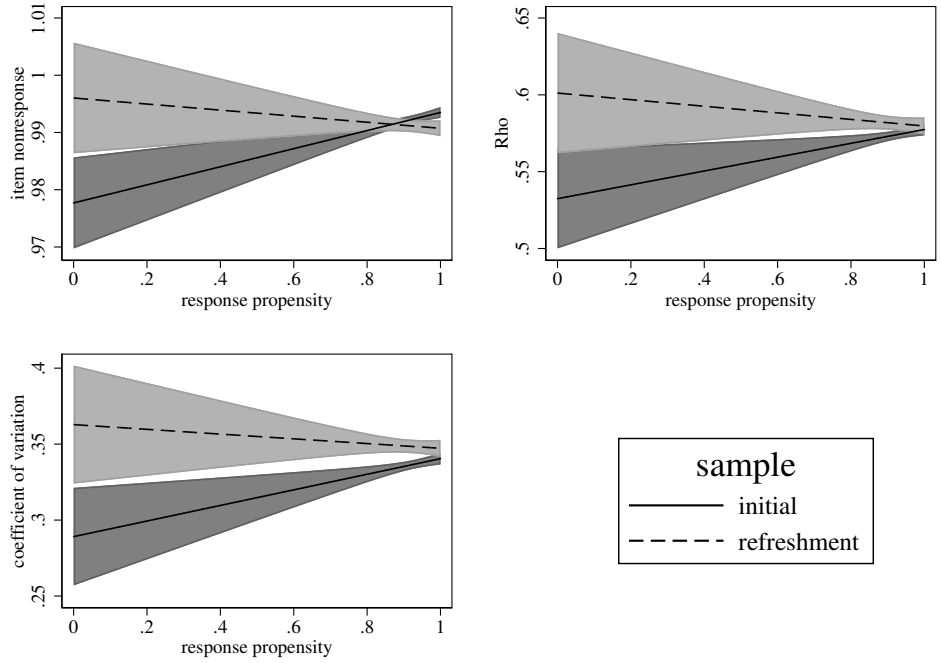

Figure 2. Relationship between response propensity and selected data quality measure

case prioritization (Peytchev, Riley, et al., 2010), may affect both nonresponse bias and measurement error. In our study, respondents with low response propensities are those most likely to produce missing answers and nondifferentiated responses. If adaptive design succeeds in increasing response by such cases, data quality may suffer. Second, response propensities could further be used to target data quality improvement interventions for high risk groups. For example, low propensity groups might be routed to a version of the questionnaire without grid questions to mitigate nondifferentiation (Roßmann et al., 2018). Third, adaptive design could specifically be designed to simultaneously reduce the risk of nonresponse and measurement errors (Calinescu & Schouten, 2016). However, the relationship between response probabilities and measurement error we find in our paper will increase the complexity of the optimization task when fielding such a design.

As always, our study can be extended and offers future research opportunities. We focus on data quality in grids because of this question type's important role in questionnaire design and its susceptibility for quality issues. With our data, we analyze the relationship between response probabilities and data quality in nine different grid questions. However, as previous research (e.g., Liu & Cernat, 2018) has shown, the design and content of a grid question may impact the response process and thus data quality. We therefore encourage future studies to extend our work and investigate a more diverse set of grid questions. Based on our experiences we would recommend to implement such a study in a long-running panel survey that features different refreshment samples. It would also be interesting if results could be compared between panels in different countries to test whether our results can be generalized beyond Germany.

In addition, our approach of estimating response propensities and using them to understand the relationship between nonresponse and data quality could be extended to other question types such as open-ended questions. Investigating further question types will require to select a different set of data quality measures that indicate issues in their specific answering process. In the case of open-ended questions, the extent and topical variety of narrative answers could be interesting to analyze (e.g., Smyth, Dillmann, Christian, & McBride, 2009).

Finally, in our study we draw on the beneficial properties of panel data. To predict response propensities we use information available from interviews in prior waves. When building similar models in cross-sectional survey, this step will be more challenging because researchers have to draw on sparse information that they are able to gather for the gross sample. If a sample is drawn from population registers most likely only basic socio-demographic information will be available (e.g., sex, age, place of residency). When continuing our line of research for cross-sectional survey, we recommend future research to devote additional attention to the estimation of response probabilities and especially the inclusion of auxiliary data that will help to improve the pre-

diction.

References

AAPOR. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. AAPOR.

Bach, R. L., Eckman, S., & Daikeler, J. (2020). Misreporting among reluctant respondents. *Journal of Survey Statistics and Methodology*, *8*(3), 566–588. doi:10.1093/jssam/smz013

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 73–93.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in germany: The GESIS panel. *Social Science Computer Review*, *36*(1), 103–115. doi:10.1177/0894439317697949

Buskirk, T. D., & Kolenikov, S. (2015). Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification. *Public Opinion Quarterly*, *74*(3), 413–432.

Calinescu, M., & Schouten, B. (2016). Adaptive survey designs for nonresponse and measurement error in multipurpose surveys. *Survey Research Methods*, *10*(1), 35–47.

Fricker, S., & Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, *74*(5), 934–955.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

GESIS Panel Team. (2018). GESIS panel - standard edition. GESIS Datenarchiv, Köln. ZA5665 Datenfile Version 24.0.0. doi:10.4232/1.13001

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Gummer, T., & Blumenstiel, J. E. (2018). Experimental evidence on reducing nonresponse bias through case prioritization: The allocation of interviewers. *Field Methods*, *30*(2), 124–139.

Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, *50*(1), 238–264.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Berlin: Springer.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley.

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, *13*(1), 73–93. doi:10.18148/srm/2019.v1i1.7395

Kreuter, F., Müller, G., & Trappmann, M. (2010). Nonresponse and measurement error in employment research: Making use of administrative data. *Public Opinion Quarterly*, *74*(5), 880–906.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*(4), 526–538.

Liu, M., & Cernat, A. (2018). Item-by-item versus matrix questions: A web survey experiment. *Social Science Computer Review*, *36*(6), 690–706.

McCarty, J. A., & Shrum, L. J. (2000). The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly*, *64*(3), 271–298.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437.

Olson, K. (2013). Do non-response follow-ups improve or reduce data quality? A review of the existing literature. *Journal of the Royal Statistical Society: Series A*, *176*(1), 129–145.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). New York, NY: Academic Press.

Peytchev, A., Peytcheva, E., & Groves, R. M. (2010). Measurement error, unit nonresponse, and self-reports of abortion experiences. *Public Opinion Quarterly*, *74*(2), 319–327.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., & Lindblad, M. (2010). Reduction of nonresponse bias through case prioritization. *Survey Research Methods*, *4*(1), 21–29.

R Core Team. (2020). *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ridgeway, G. (2017). Gbm: Generalized boosted regression models. *R package version 2.1.3*. https://cran.r-project.org/package=gbm.

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

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*, *6*(3), 376–400.

Schouten, B., Cobben, F., Lundquist, P., & Wagner, J. (2016). Does more balanced survey response imply less nonresponse bias? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*(3), 727–748.

Schouten, B., Peytchev, A., & Wagner, J. (2017). *Adaptive survey design*. CRC Press.

Smyth, J. D., Dillmann, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*(2), 325–337.

Tourangeau, R. (2019). How errors cumulate: Two examples. *Journal of Survey Statistics and Methodology*, *(online first)*.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368–393.

Tourangeau, R., Maitland, A., Rivero, G., Sun, H., Williams, D., & Yan, T. (2017). Web surveys by smartphone and tablets: Effects on survey responses. *Public Opinion Quarterly*, *81*(4), 896–929.

Tourangeau, R., Michael Brick, J., Lohr, S., & Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society: Series A*, *180*(1), 203–223.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Wagner, J. (2008). *Adaptive survey design to reduce nonresponse bias* (Doctoral dissertation).

Wagner, J., West, B. T., Kirgis, N., Lepkowski, J. M., Axinn, W. G., & Ndiaye, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, *28*(4), 477–499.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35.

Appendix A
Figure

(12) Inwieweit treffen die folgenden Aussagen auf Sie persönlich zu?

	trifft gar nicht zu	trifft eher nicht zu	trifft teilweise zu	trifft eher zu	trifft voll und ganz zu
Ich bin gut darin, Versuchungen zu widerstehen.	o	o	o	o	o
Es fällt mir schwer, schlechte Gewohnheiten abzulegen.	o	o	o	o	o
Ich bin faul.	o	o	o	o	o
Ich sage unangemessene Dinge.	o	o	o	o	o
Ich tue manchmal Dinge, die schlecht für mich sind, wenn sie mir Spaß machen.	o	o	o	o	o
Ich wünschte, ich hätte mehr Selbstdisziplin.	o	o	o	o	o
Angenehme Aktivitäten und Vergnügen hindern mich manchmal daran, meine Arbeit zu machen.	o	o	o	o	o

Figure A1. Screenshot of question grid on personality in GESIS Panel wave ed.

Appendix B
Tables

Table B1
Predictor variables used in nonresponse model

Predictor variables
Trust: general trust
Private internet usage
Frequency private Internet usage: PC
Private internet usage: smart phone
Private internet usage: tablet PC
Survey experiences in total
Survey experiences online
Survey experiences postal
Survey experiences personal
Gender
Country of birth mother, region 3 categories
Marital status, 5 categories
Steady partner
Joint household
Highest school leaving certificate, incl. N/A
Vocational or professional training, incl. N/A
Employment situation
Number of children under 16 years, 3 categories
Personal income, 15 categories
Household income, 14 categories
Television consumption
Radio consumption
Newspaper consumption
Most important problem in Germany
Second most important problem in Germany
Satisfaction federal government
Satisfaction democracy in Germany
Political interest
Left-right-scale
Civic duty: Social/ political activity
Civic duty: Political consume
Civic duty: Military service
Quality of life in region
Affected by environmental influences: Noise pollution
Affected by environmental influences: Air pollution
Affected by environmental influences: Missing accessible public parks
Social relationship neighborhood
Previously changed place of residence
More time for: Read books
More time for: Gardening
More time for: Home improvement

Continues on next page

Continued from last page

More time for: Watching TV
More time for: Doing sports
More time for: Going out
More time for: Travel
More time for: Surf the internet/ play on the computer
More time for: Children/ grandchildren
More time for: Partner
More time for: Relatives/ friends
More time for: Voluntary activities
More time for: Hobbies
More time for: Go shopping
More time for: Others (indicator dichotomous)
More time for them: Other - open ended
Personal priority: Be able to afford something
Personal priority: Be there for others
Personal priority: Fulfil oneself
Personal priority: Successful career
Personal priority: Own house
Personal priority: Fortunate marriage/partnership
Personal priority: Children
Personal priority: Campaign politically/ socially
Personal priority: See the world, make a lot of voyages
Mode of invitation
Mode of participation
Mode of invitation at first wave
AAPOR wave code
Survey Evaluation: Interesting
Survey Evaluation: Diverse
Survey Evaluation: Important for science
Survey Evaluation: Long
Survey Evaluation: Difficult
Survey Evaluation: Too personal
Overall assessment
Participation interrupted
Other people present during interview
Participation Location
Participation device
Year of birth, extreme values summarized
Participation history: Counting response in the previous waves

Table B2
Regression estimates on relationship between response propensities and data quality measures

	Left-aligned responses b (se)	Extreme responses b (se)	Item nonresponse b (se)	Straightlining b (se)	ρ b (se)	CV b (se)	Failing attention check b (se)
Response propensity	−0.026	−0.033	0.016***	1.420*	0.045*	0.051**	−0.259
	(0.016)	(0.032)	(0.004)	(0.686)	(0.017)	(0.018)	(0.862)
Sample: Initial	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Refreshment	−0.047*	−0.107*	0.018**	1.499	0.069**	0.074**	−0.689
	(0.023)	(0.047)	(0.006)	(1.087)	(0.026)	(0.026)	(1.184)
Refreshment ×	0.043	0.090	−0.021**	−1.434	−0.066*	−0.067*	0.811
Response propensity	(0.025)	(0.051)	(0.007)	(1.172)	(0.028)	(0.028)	(1.275)
N	38313	38313	38313	38313	36629	36629	2320
Log Likelihood	23442.198	5633.060	57841.619	-4203.463	20061.247	21205.882	-985.557

Intercept and controls omitted from output.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$