mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology

Volume 19, 2025 | 1

Matthias Sand et al. — Creating Design Weights for a Panel Survey With Multiple Refreshment Samples: A General Discussion With an Application to a Probability-Based Mixed-Mode Panel

Frederick G. Conrad et al. — Probing in Cognitive Interviews Can Promote Acquiescence

Florian Heinritz — Mother Tongue or Non-Native Language? – The Influence of Language on Response Behavior in Surveys

Jessica Donzowa et al. — From Clicks to Quality: Assessing Advertisement Design's Impact on Social Media Survey Response Quality

Maura Spiegelman & Frederick Conrad — Improving Understanding of Survey Questions with Multimodal Clarification

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

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.

Copyediting: Désirée Nießen, Barbara Felderer, Maximilian Linde
Layout: Liebchen + Liebchen
ISSN 2190-4936 (Online)

Content

RESEARCH REPORTS

4 Creating Design Weights for a Panel Survey With Multiple Refreshment Samples: A General Discussion With an Application to a Probability-Based Mixed-Mode Panel
Matthias Sand, Christian Bruch, Barbara Felderer, Ines Schaurer, Jan-Philipp Kolb & Kai Weyandt

23 Probing in Cognitive Interviews Can Promote Acquiescence
Frederick G. Conrad, Rachel E. Davis, Carolyn Lau, Melissa Armendáriz, Stephanie Morales, Timothy P. Johnson & Johnny Blair

67 Mother Tongue or Non-Native Language? – The Influence of Language on Response Behavior in Surveys
Florian Heinritz

94 From Clicks to Quality: Assessing Advertisement Design's Impact on Social Media Survey Response Quality
Jessica Donzowa, Simon Kühne & Zaza Zindel

137 Improving Understanding of Survey Questions with Multimodal Clarification
Maura Spiegelman & Frederick Conrad

Creating Design Weights for a Panel Survey With Multiple Refreshment Samples: A General Discussion With an Application to a Probability-Based Mixed-Mode Panel

Matthias Sand[1], Christian Bruch[1], Barbara Felderer[1], Ines Schaurer[2], Jan-Philipp Kolb[3] & Kai Weyandt[1]

[1] *GESIS – Leibniz Institute for the Social Sciences*
[2] *City of Mannheim*
[3] *Federal Statistical Office of Germany (DESTATIS)*

Abstract

Panel surveys suffer from attrition, where participants drop out over time. To maintain generalizability, refreshment samples are frequently employed, bringing in new individuals, increasing the number of panelists, and balancing sample composition. Although refreshment samples offer numerous advantages, the inclusion of new panel members may introduce bias into the analysis if the design weights are not appropriately tailored to these new members and adjusted to align with existing panel members. If not correctly accounted for, their inclusion may bias results. This paper addresses the issue of designing proper weights by applying the multiple-frame weighting approach proposed by Kalton and Anderson, which is generally used for cross-sectional surveys, to ongoing panel studies with refreshment samples. We demonstrate its application to a synthetic data set and a probability-based mixed-mode panel with an initial sample and two refreshment samples. We compare estimates obtained using multiple-frame weighting with those obtained using unweighted and naively weighted methods (where design weights are used as calculated for the respective samples without adjusting for the fact that some members of the population have a chance of being sampled more than once due to the refreshments). These comparisons showcase the potential for bias introduced by neglecting proper weighting and underscore the importance of both a multiple-frame weighting approach and meticulous sample documentation.

Keywords: panel surveys, GESIS Panel, refreshment samples, multiple-frame weighting, inclusion probabilities

To study social change, panel surveys of the same individuals over time are crucial. Ensuring the validity of the panel's findings requires that the panel members adequately represent the population. Panel surveys, which usually start with a random sample drawn from the population of interest, face attrition, as some panel members choose to discontinue their participation, can no longer be contacted, or die. Attrition introduces the risk of a panel being selective for certain population subgroups, especially if members of some subgroups drop out at higher rates than others. In addition to potential attrition bias, the reduced sample size decreases the precision of sample estimates.

To counteract the negative effects of attrition, panels such as the Longitudinal Internet Studies for the Social Sciences (LISS) panel (Scherpenzeel, 2011), the German Internet Panel (Blom, Gathmann, & Krieger, 2015), and the GESIS Panel (Bosnjak et al., 2018) are usually refreshed after some time by recruiting new panel members. In scientific research, both the initial recruitment sample and the refreshment sample(s) are usually drawn using a random sampling approach. It may be a simple task to determine for each recruitment sample an individual's propensity to be sampled (in the case of sampling designs that are not too complex). However, combining the initial recruitment sample and refreshment samples drawn at different points in time from a population of interest that is described in the same way (e.g., persons aged 18 years or older) is not a trivial task. This is due to the fact that each sample is drawn sequentially and independently of the previous samples. One key challenge is therefore to account for the fact that, in principle, some members of the population have a chance of being sampled more than once, whereas others do not, as they were not part of the population of interest when the previous samples were recruited. This results in very different probabilities of being included in the panel survey. Naive weighting strategies, such as directly adopting design weights using the design weights of the individual samples without adjusting for potential overlap between the sampling frames or the probability of being sampled several times, fail to yield valid inferences for panel surveys with refreshments, as they would lead to an overestimation of the population in cases where the population of interest of each recruitment overlaps (Gabler et al., 2012; Lohr, 2011; Sand & Gabler, 2018).

In this paper, we show how the multiple-frame weighting methodology proposed by Kalton and Anderson (1986) and Lohr and Rao (2006), which was originally developed for cross-sectional surveys with more than one sampling frame, can be used to create weights in a panel context. We demonstrate that using the initial design weights for the recruitment and refreshment samples

Direct correspondence to
Matthias Sand, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
E-mail: Matthias.Sand@gesis.org

separately may result in significant biases, even when these samples are individually self-weighted. However, there is a limited body of literature on the correct calculation of design weights for panel surveys with refreshment samples. Therefore, using the GESIS Panel—a German probability-based mixed-mode panel—as an illustrative example, we showcase how the multiple-frame weighting approach provides more accurate estimates. We assess the weights by comparing unweighted, naively weighted, and multiple-frame-weighted estimates of age and region with their corresponding actual population values.

The remainder of the paper is structured as follows: In the next section (Multiple-Frame Approach), we introduce the multiple-frame weighting approach proposed by Kalton and Anderson (1986) and discuss how it can be applied and understood in the panel context rather than its original context of application, multiple cross-sectional samples. We further demonstrate its use under ideal (and controlled) conditions using a synthetic data set. In the third section (Applying the Multiple-Frame Approach to the GESIS Panel), we apply this approach to actual panel data. We conclude with a discussion of our findings (Discussion).

Multiple-Frame Approach

Sampling theory allows for the use of inclusion probabilities from the sampling design to estimate population values (e.g., totals) with the Horvitz-Thompson estimator (Horvitz & Thompson, 1952). However, when a survey is conducted using several sampling frames that partially cover the entire population, frames may intersect. Therefore, multiple frames, which are common in real-world surveys, require multiple-frame approaches to address the overlap of frames and ensure unbiased estimates and/or to calculate the inclusion probabilities. Approaches such as that proposed by Kalton and Anderson (1986) adjust the inclusion probabilities to account for individuals appearing in multiple frames by incorporating the overlap into the estimation process. Therefore, such an approach prevents overestimation of the accessible population (Brick et al., 2005; Lohr & Rao, 2006; Sand & Gabler, 2018; Skinner & Rao, 1996).

A common application of such an approach is a telephone survey of all residents in a country. In such surveys, two sampling frames are typically used: a list of all landline numbers and a list of all mobile phone numbers. Neither list contains all or most of the population members; for example, younger individuals may be missing from the landline list, and older individuals may be missing from the mobile phone list (Heckel & Wiese, 2011). When conducting surveys using two different sampling frames, two types of individuals can be identified: those who can participate in the survey via both frames and those who can participate only via one of the frames.

The challenge lies in the potential overlap of the different sampling frames, as some individuals may be accessible via both frames, thereby increasing their likelihood of being selected for the survey. This circumstance must be accounted for by using a multiple-frame approach when calculating design weights. Several methods can be used during the estimation process to account for individuals being part of multiple sampling frames. The most notable methods (e.g., the multiplicity approach or convex combinations) involve transforming or weighting the design weights of individuals belonging to both frames (Brick et al., 2005; Singh & Mecatti, 2011) or calculating a joint inclusion probability, as in the Kalton–Anderson approach (Kalton & Anderson, 1986; Lohr, 2007).

Multiple-frame approaches are commonly used in cross-sectional surveys. We propose to view the initial sample and the refreshment samples in panel surveys as multiple frames and to apply a multiple-frame approach to derive accurate design weights. In panel surveys, where the same group of individuals is surveyed repeatedly over time, fluctuations due, for example, to migration, births/deaths, or aging in the population may also pose a challenge. As the composition of the population changes, some individuals may become unreachable or no longer meet the survey's eligibility criteria. To address this issue, researchers may opt to use refreshment samples, which involve introducing new participants into the panel to replace those who have attrited or become ineligible. Additionally, the population from which the initial sample was drawn also ages. Hence, when initiating a refreshment, a compelling case can be made for treating the dynamic fluctuations within a population from one point in time to another as distinct frames, albeit with a substantial overlap. Recognizing these temporal shifts as separate frames is crucial to prevent biased estimates, particularly when there is a risk of overestimating subpopulations sampled at multiple time points. To address this concern, adopting a multiple-frame approach becomes imperative. It is noteworthy that the existing literature focuses predominantly on cross-sectional surveys and that there is a notable dearth of documented applications of multiple-frame approaches to panel surveys. This underscores the urgency of considering the temporal evolution within a population as distinct frames—especially when conducting refreshments—to ensure more accurate and unbiased estimates.

Panels With Refreshment Sample(s) as a Special Case of a Multiple-Frame Survey

Assuming a panel survey originates from a simple random sample design and incorporates refreshment sample(s) at a specific point in time to counteract the adverse effects of panel attrition, it is essential to consider changes in the inclusion probabilities for each element in the panel survey. If the sampling frame for the refreshment sample(s) has not been adjusted for the elements that were

already part of the original (gross) sample, there is a theoretical possibility of an element drawn in the initial sample also entering the second sample. Furthermore, the inclusion probabilities of these specific elements increase with each subsequent refreshment. Additionally, due to the assumed time gap between the original sample and the refreshment sample, the latter may include elements that were not yet eligible at the time the original sample was drawn and that entered the panel exclusively via the refreshment sample. Consequently, regarding the estimation based on this "refreshed" panel survey, it can be viewed as a multiple-frame survey, given the disparity between the original and refreshment sampling frames, even though considerable overlap is presumed.

The refreshment of a panel survey may take place at several points in time. The different sampling frames may occur similarly to the Venn diagram in Figure 1.

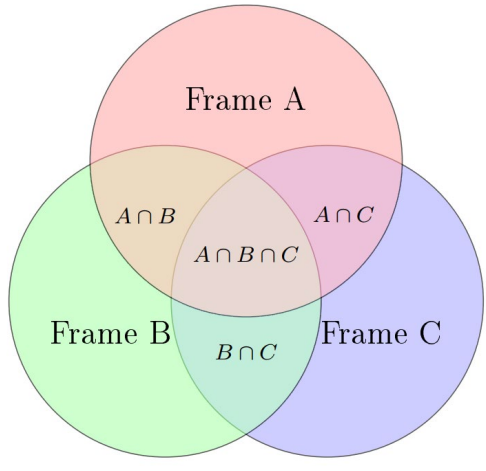

Figure 1 Schematics of a multiple-frame survey comprising three sampling frames

In the context of panel surveys, we have the original sample drawn from Frame A and two refreshment samples drawn from Frames B and C (see Figure 1). As the refreshment samples are generally drawn sequentially at two different points in time, we end up with three different frames, two or more of which overlap. In the case of a panel survey with several refreshments, one might assume the intersections between the frames to be considerable.

This implies that individuals who are in two or three sampling frames (and therefore form the intersection) could enter the panel at multiple time points and thus have a higher probability of being sampled for the panel compared

with those who are part of only one of the frames. For individuals within the intersection of two or more frames, it is crucial to accurately calculate the inclusion probability, accounting for the possibility of being sampled from more than one frame, to prevent estimation bias.

Simply using the inclusion probability of each of the three samples based on each of the sampling frames may lead to an overestimation of the number of elements within the intersections and an underestimation of those that can be sampled from only one frame. However, simply adding up the probabilities of inclusion for those elements within each intersection would lead to an overestimation of their inclusion probability. Therefore, it is crucial to accurately calculate an individual's overall inclusion probability by appropriately adding and subtracting their corresponding joint inclusion probabilities of each frame. This mechanism includes and excludes particular overlaps of the respective frames when calculating the inclusion probability. In this particular example, the design weights must be generated as follows: For each of the three samples, the probability of being included in the corresponding sample s_h, with $h \in A, B, C$ is given by

$$\pi_i^{s_h} = \frac{n^{s_h}}{N^{s_h}}, \tag{1}$$

where n refers to the (gross) sample size of a sample and N to the number of elements within each sampling frame.

To adjust for the multiple-frame sampling design, three groups of individuals can be distinguished: (a) individuals who can enter the survey via all three sampling frames, (b) individuals who can enter the survey via two sample frames, and (c) individuals who can enter the survey via only one of the three sampling frames.

For individuals in Group (a), the (adjusted) inclusion probabilities are given by

$$\begin{aligned}\pi_i = (\pi_i^{s_A} + \pi_i^{s_B} + \pi_i^{s_C}) - (\pi_i^{s_A} * \pi_i^{s_B}) - (\pi_i^{s_A} * \pi_i^{s_C}) - (\pi_i^{s_B} * \pi_i^{s_C}) + \\ (\pi_i^{s_A} * \pi_i^{s_B} * \pi_i^{s_C}).\end{aligned} \tag{2}$$

Group (b) consists of three subgroups: (1) individuals who can enter the survey via Samples A and B, (2) individuals who can enter it via Samples A and C, and (3) individuals who can enter it via Samples B and C. For the first subgroup, π_i could be derived by setting the inclusion probability of the frame of which the individuals are not part in equation (2) to zero. The inclusion probabilities are then given by

$$\pi_i = (\pi_i^{s_A} + \pi_i^{s_B}) - (\pi_i^{s_A} * \pi_i^{s_B}) \tag{3}$$

Inclusion probabilities for the other two subgroups are generated accordingly. For individuals in Group (c), the multiple-frame inclusion probabilities are equal to the inclusion probabilities of the corresponding sample.

In the present paper, the importance of using the correct design weights to perform inference from a multiple-frame survey will be illustrated in a synthetic data example and the GESIS Panel (Bosnjak et al., 2018). There are other methods for adjusting design weights in multiple-frame designs, for example, fixed weight adjustment by Hartley (1962), the multiplicity approach by Mecatti (2007), and the pseudo-maximum likelihood method by Lohr and Rao (2006). Sand (2018) showed that using a composite approach can lead to more precise estimates, while point estimates are almost identical to the Kalton-Anderson approach. Composite weighting approaches adjust the design weight of an element in the overlap population by a factor between 0 and 1. However, these approaches require further information on the sampling frames, their respective sizes and overlap, or knowledge about the original frame that was used to sample a particular element. In our case, using only the data set of the GESIS Panel that was provided to us, we could not identify the original sampling frame. However, as we worked only on a reduced GESIS panel data set that included only age and region, it was easy to determine whether an individual belonged to the overlap population. Using the full GESIS Panel data set, one could nevertheless also employ a composite approach similar to that suggested by Brick et al. (2005).

Illustration of the Multiple-Frame Approach Based on a Synthetic Data Set

To demonstrate the workings of a multiple-frame weighting approach for panels with refreshment samples under controllable conditions, we initially generated a synthetic data set, mimicking the sampling approach and the related sampling frames of the GESIS Panel. The synthetic population was constructed in accordance with official statistics.

In our example, we assumed that the recruitment of the original sample started with a population aged between 18 and 69 years (Frame 1). Two years later, a first refreshment sample was drawn. Hence, each member of the synthetic population who was at least 16 years old when the initial panel recruitment started (and 18 years old at the time of the first refreshment) could be part of that refreshment sample (Frame 2). Three years after that, the second refreshment sample was drawn in accordance with the design of the first refreshment (Frame 3). For that particular sample, each individual who was at least 13 years old when the initial panel recruitment took place could be part of the second refreshment (Frame 3). All three frames together cover a population of 68 million elements (100%). Frames 1 and 2 jointly cover 65.7 million elements (96.6%), and Frame 1 contains only 53.4 million elements (78.5%). Table 1 illustrates the

varying target populations of the underlying sampling frames based on the age of the persons at the time of recruitment of the original sample.

Table 1 Target populations of the underlying sampling frames based on the age of the persons at the time of recruitment of the original sample

Age category[1]	Frame 1	Frame 2	Frame 3
13–15 years	✗	✗	✓
16–17 years	✗	✓	✓
18–69 years	✓	✓	✓
70+ years	✗	✓	✓

[1] The age category refers to an individual's age at the time of recruitment of the original sample.

As can be seen in Table 1, there is an overlap of all three sampling frames for those elements aged 18–69 years when the original sample was recruited, and there is an overlap of Frames 2 and 3 for those aged 16–17 years and 70 years and over when the original sample was recruited. However, those aged 13–15 years when the original sample was drawn could come only from Frame 3. As this simulation study mimics the approach of the GESIS Panel, the recruitment of the first and second refreshments (Frames 2 and 3) differs from that of the initial sample (Frame 1). Whereas the initial sample was restricted by a maximum age of 69, the first and second refreshments were not. Hence, persons who were at least 70 years old when the original sample was recruited could be sampled only from Frames 2 and 3.

Similar to sampling designs often used in Germany, we further divided the synthetic population into two strata, "east" and "west," in accordance with the distribution of the population across eastern and western German federal states. We did so to achieve a close approximation of the GESIS Panel, which will be discussed in the next section.

From the synthetic population divided into the strata "east" and "west," we drew the three samples using an approach similar to that used by the GESIS Panel. We also used the GESIS Panel's gross sample sizes (see Table 2).

As can be seen in the last column of Table 2, the sample size of the initial sample was allocated proportionally to both strata, whereas the sample sizes of the two refreshment samples were disproportionally allocated, with an oversampling of elements stemming from the stratum "east."

Table 2 Sample sizes and allocation of the sample sizes in the synthetic data set to the strata "east" and "west"

Sample	Stratum	No. of elements	Proportion of sample
Frame 1		**21,870**	**100%**
	East	3,716	16.99%
	West	18,154	83.01%
Frame 2		**10,692**	**100%**
	East	3,366	31.48%
	West	7,326	68.52%
Frame 3		**11,502**	**100%**
	East	3,621	31.48%
	West	7,881	68.52%
Total size		**44,054**	

Let us now assume that we want to estimate the age distribution of the population based on the survey data. We can apply three strategies:

1. Use the unweighted estimates to infer the population.
2. Apply a naive weighting approach by using the design weights of the indivdual samples without adjusting for potential overlap between the frames. Design weights would then be based on the inclusion probability $\pi_i^{s_A}$ for individuals who were sampled during the initial recruitment and on the inclusion probabilities $\pi_i^{s_B}$ and $\pi_i^{s_C}$ for individuals who were sampled in the first and second refreshments, respectively.
3. Apply design weights generated according to the multiple-frame approach described above.

The first two strategies are considered here due to their potential misuse when analysts are unaware of the multiple-frame approach or the issues discussed in the section entitled "Panels With Refreshment Samples as a Special Case of a Multi-Frame Survey." These strategies can be applied when design weights are not provided or are available only on request (e.g., the LISS panel; see https://www.lissdata.nl/faq). Misapplication may also occur if the panel provider publishes incorrect design weights, as noted by Wetzel, Schumann, and Schmiedeberg (2021) in their correction for the pairfam panel. Our objective in exploring these approaches was to highlight their adverse impacts and underscore the necessity of adopting the multi-frame approach. We therefore decided to forgo any further adjustments of these weights (e.g., for nonresponse or panel attrition).

An initial evaluation of the accuracy of a design-weighted estimator involves cross-referencing the sum of (unscaled) design weights with the actual population size. As mentioned earlier, the population of all three frames together comprises 68 million elements. With the multiple-frame approach, the sum of the design weights was 67.99 million, whereas the naive approach—which does not account for the overlap between frames or the possibility of being sampled several times—yielded a total of 186.97 million. Both estimates refer to the full set of the panel's population at the second refreshment. This stark contrast makes it evident that the naive approach would substantially overestimate the population size, a consequence of the issues discussed in the preceding section. Table 3 shows the resulting estimations for the age distribution—for example, the respective percentages of population members who belonged to the 10 age categories—applying the three different weighting approaches to the original sample and the two refreshments.

Table 3 Example: Estimation of age with the synthetic data set using the three different weighting approaches

Age category	Unweighted estimation	Naive estimation	Multiple-frame estimation	True population value
13–15 years	0.79	1.16	3.18	3.38
16–17 years	1.10	1.67	2.29	2.18
18–29 years	18.41	17.66	16.21	16.04
30–39 years	15.47	14.88	13.64	13.36
40–49 years	21.07	20.11	18.50	18.79
50–59 years	18.76	17.55	16.22	16.37
60–69 years	14.45	13.80	12.54	5.16
70–79 years	6.87	8.84	11.46	11.45
80–89 years	2.67	3.74	5.14	5.16
90+ years	0.41	0.60	0.82	0.78

Comparing the estimates obtained using the three strategies with the true population values (last column), one can easily see that the multiple-frame estimates are closest to the population values and that differences are likely attributable to sampling error alone. The estimates obtained using the unweighted and naive approaches are particularly poor for the youngest and oldest age cohorts. This shows that they cannot account for the fact that individuals in these age cohorts could be sampled by only one or both of the refreshment samples, whereas individuals in the overlapping cohorts could, in addition, be sampled in the initial

recruitment. Similar but less pronounced effects were found when estimating the east/west distribution, as can be seen in Table 4.

Table 4 Example: Estimation of east/west distribution with the synthetic data set

Stratum	Unweighted estimation	Naive estimation	Multiple-frame estimation	True population value
East	24.29	16.02	15.99	15.95
West	75.71	83.98	84.01	84.05

Comparing the unweighted and naive approaches, one can see that the naive approach performed much better. This is due to the fact that the design weights for the refreshment samples accounted for the oversampling of eastern Germany. The smaller discrepancy between naive and multiple-frame estimation in the case of the east/west estimation might be explained by the similar distribution of the strata across all age classes.

To conclude, using unweighted or naively design-weighted estimations misrepresents the age distribution and might severely bias population inference. As described in the next section, to test these results on an actual panel, we applied these weighting strategies to the GESIS Panel.

Applying the Multiple-Frame Approach to the GESIS Panel

The GESIS Panel, a mixed-mode panel representing the general population in Germany (Bosnjak et al., 2018), employs self-administered surveys conducted bimonthly until 2020 and every third month from 2021 onward. The initial recruitment process involved a multi-step transition from interviewer-administered personal recruitment interviews to self-administered surveys. The survey initially targeted persons living in Germany aged between 18 and 69. Two refreshment samples were drawn in 2016 and 2018 using the German General Social Survey (ALLBUS) interview as the recruitment interview. The ALLBUS applies a two-stage sampling process, but it oversamples the region of eastern Germany. The target population differs from the initial sample and is defined as persons older than 17 years living in Germany, without an upper age limit. Table 5 displays each GESIS Panel sample, its respective design, and its target population.

Table 5 Design and target population of the GESIS Panel recruitment samples

Sample	Age range	Sampling approach
2013 initial cohort	18–70 years	Self-weighted
2016 refreshment (*R*1)	17 years and older	Oversampling of eastern Germany; self-weighted within stratum
2018 refreshment (*R*1)	17 years and older	Oversampling of eastern Germany; self-weighted within stratum

Figure 2 illustrates the target populations of the various GESIS Panel samples, emphasizing changes in eligibility criteria between the initial cohort and the refreshment samples. Notably, individuals born before 1942 and a younger age cohort are included in refreshment samples, thereby expanding the panel's coverage.

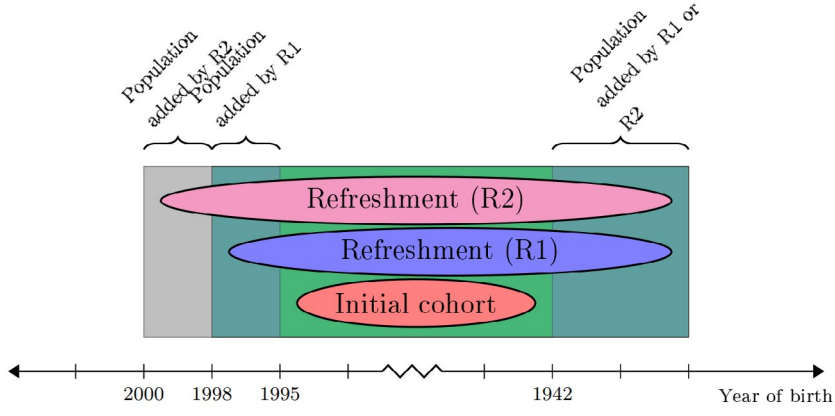

Figure 2 Target population of the initial cohort and the refreshment samples of the GESIS Panel

Deriving Design Weights of the GESIS Panel

When combining the initial cohort of the GESIS Panel with its two refreshment samples, several points must be considered. First, each sample was drawn at different times, leading to slightly different target populations. Second, the refreshment samples had different age restrictions compared with the initial cohort. Thus, individuals could potentially have been included in one, two, or all

three of the GESIS Panel samples. Finally, the design weights must account for the disproportional allocation of sample size to eastern and western Germany in the first and second refreshment samples.

As already discussed, the initial cohort stems from a self-weighted sampling design. Hence, each element of the initial cohort (IC) has the same inclusion probability π_i^{IC} given by:

$$\pi_i^{IC} = \frac{n^{IC}}{N^{IC}} \tag{4}$$

In the second and third recruitment ($R1$, $R2$) the design weighting has to compensate for the disproportional allocation of sample size between eastern and western Germany in the ALLBUS sampling design. Thus, weights must be calculated separately for eastern and western Germany (GESIS, 2021). With $k \in \{East, West\}$ being an indicator for western or eastern Germany, inclusion probabilities are given by

$$\pi_{i,k}^{R1} = \frac{n_k^{R1}}{N_K^{R1}} \tag{5}$$

and

$$\pi_{i,k}^{R2} = \frac{n_k^{R2}}{N_K^{R2}}. \tag{6}$$

As described in the section entitled “Multiple-Frame Approach,” the GESIS Panel can be regarded as a three-frame design with its initial recruitment and two refreshments. A sizeable overlap of the three frames can be observed. Individuals who were born between December 1, 1942, and November 30, 1995, could—at least theoretically—have been sampled at each of the three recruitments. Hence, the inclusion probability must be adjusted in accordance with Equation (2). Moreover, due to the disproportional allocation of sample size to eastern and western Germany in Refreshments 1 and 2, an individual’s inclusion probability can be written as

$$\pi_{i,k} = \left(\pi_{i,k}^{R2} + \pi_{i,k}^{R1} + \pi_i^{IC}\right) - \left(\pi_{i,k}^{R2} * \pi_{i,k}^{R1}\right) - \left(\pi_{i,k}^{R2} * \pi_i^{IC}\right) - \left(\pi_{i,k}^{R1} * \pi_i^{IC}\right) + \left(\pi_{i,k}^{R2} * \pi_{i,k}^{R1} * \pi_i^{IC}\right). \tag{7}$$

Persons born before December 1942 and persons born between December 1995 and November 1998 could be recruited only in the first and second refreshments. In their case, π_i^{IC} would be zero. For persons born between December 1998 and November 2000, the equation above reduces to $\pi_{i,k}^{R2}$.

Comparison of Weighting Strategies in the GESIS Panel

In this section, we describe the application of multiple-frame weighting to the actual data from the GESIS Panel. Similar to the synthetic data example presented in the section entitled "Illustration of the Multiple-Frame Approach Based on Synthetic Data," we conducted comparisons with the unweighted and naive estimations. First, we assessed the population size estimates. Table 6 presents the design weights, gross sample sizes, and estimated population sizes for the multiple-frame weighting approach. Table 7 provides the same estimates using the naive weighting approach.

As can be seen in Table 7, the naive weighting approach yielded an overall population estimate of 192.5 million, significantly surpassing the actual population size of about 68.8 million. This overestimation stems from the naive method of iteratively calculating population sizes, resulting in inflated figures due to the repeated consideration of the intersection. When applying design weights without adjustments, the extrapolated population size equaled the sum of the three samples, whereas focusing, for example, solely on the elements and weights of the second refreshment ($R2$) yielded the correct population size of 68.85 million (56.95 million + 11.9 million). Consequently, the design weights of each sample extrapolated to the corresponding sampling frame's population size.

By contrast, the multiple-frame approach estimated an overall population size of 69.2 million (Table 6), exhibiting a slight overestimation compared with the true population size. This discrepancy may have arisen from various limitations in the real data, such as incomplete control of the population and a lack of knowledge regarding the distribution of relevant variables. Additionally, errors in element specification, reporting incorrect population information, and discrepancies in frame and gross sample sizes, coupled with the inability to retrospectively track each step of the sampling process, contributed to more pronounced discrepancies compared with the synthetic data examples. Despite the demonstrated accuracy of the multiple-frame approach under ideal conditions, these factors appear to have influenced its results in this real-data scenario. Furthermore, the observed difference might be attributable to individuals appearing in multiple groups, a possibility that cannot be ruled out due to the absence of detailed information on individual appearances across groups. We further compared the distribution of age (Table 8) and region (Table 9) in a similar way as we did for the synthetic data set.

Table 6 Distribution of weights and population size estimation with the multiple-frame approach by different age cohorts, separately for eastern and western Germany

Category	Weight	Gross sample size	Estimated population size
Born before 12/1942, west	3,710.73	2,065	7,662,658
Born before 12/1942, east	1,700.12	1,129	1,919,438
Born between 12/1942 and 11/1995, west	1,514.51	30,610	46,359,138
Born between 12/1942 and 11/1995, east	1,021.60	9,372	9,574,408
Born between 12/1995 and 11/1998, west	3,710.73	484	1,795,994
Born between 12/1995 and 11/1998, east	1,700.12	144	244,818
Born after 11/1998, west	7,226.69	202	1,459,791
Born after 11/1998, east	3,285.41	58	190,554
Overall population			**69,206,798**

Table 7 Distribution of weights and population size estimation with the naive approach by different cohorts

Category	Weight	Gross sample size	Estimated population size
Initial cohort	2,558.223	21,870	55,948,331
Refreshment (*R*1), west	7,625.969	7326	55,867,847
Refreshment (*R*1), east	3,522.321	3,366	11,856,132
Refreshment (*R*2), west	7,226.688	7,881	56,953,526
Refreshment (*R*2), east	3,285.413	3,621	11,896,481
Overall population			**192,522,317**

Table 8 Weighted distribution of age group (in %)

Age class	Unweighted estimation	Naïve estimation	Multiple-frame estimation	True population value
Born before 1943	7.25	9.96	13.85	12.71
Born between 01/1943 and 12/1995	90.74	87.05	80.82	82.02
Born between 01/1996 and 12/1998	1.43	2.13	2.95	4.02
Born after 12/1998	0.59	0.86	2.38	1.26

Table 9 Weighted distribution of region (in %)

Region	Unweighted estimation	Naïve estimation	Multiple-frame estimation	True population value
West	75.71	82.72	82.76	82.72
East	24.29	17.28	17.24	17.28

We used the 2018 German intercensal population updates ("Fortschreibung des Bevölkerungsstandes;" Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, 2021) as our source of official statistics; they also served as the sampling frame for drawing samples constituting the three cohorts. Derived from Germany's 2011 census data, this model-based estimation includes statistical uncertainty. The multiple-frame weighting approach provided the closest estimation of the true population value for most age categories, except for the youngest age group. A crucial factor contributing to this accuracy is that the multiple-frame method takes into consideration the exclusion of certain age categories in specific recruitment waves, a nuance overlooked by the naive estimation. For instance, the initial cohort does not include the oldest respondents and the second-youngest age group, thus distorting the inter-category relations when combining all cohorts. Nevertheless, the naive estimation remained closer to the population value than did the unweighted estimation.

Regarding the variable "region," both naive and multiple-frame estimations aligned closely with the true population value, which was anticipated, as this variable influences the weighting across all cohorts. By contrast, the unweighted estimation significantly deviated from the population benchmark due to the disproportional allocation of sample size to eastern and western Germany in the ALLBUS.

However, the analyses presented in this section have certain limitations. The absence of information on pertinent frame parameters, particularly concerning the population and frame distribution of the former East and West Berlin, poses challenges. We assume that during the refreshment sampling, the population of the former East Berlin was part of the "east" stratum, thereby leading to oversampling. Determining participants' ages accurately at the time of the refreshments was also problematic, and retrospective reconstruction of this specific aspect of the frame proved unfeasible.

Discussion

In this paper, we examined how refreshment samples can be integrated correctly into panel surveys using the multiple-frame approach. The differences between multiple-frame weighting and a naive weighting approach were illustrated using a synthetic data set. We show that the estimates using multiple-frame weighting deviated only slightly and at random from the population parameters, whereas naively weighted and unweighted estimates showed large systematic discrepancies. Applying the approach to the real data of the GESIS Panel, we found the differences between the naive weighting procedure and the multiple-frame approach to be less pronounced.

The inability to fully replicate the findings from our synthetic data set when using actual panel data can be attributed to issues arising from the time gap between the calculation of weights and the sampling conducted by a third-party field agency. To achieve accurate weights, comprehensive information about the sampling process and the data used to design the survey sample is imperative. Any uncertainties or discrepancies in this information pose a potential risk to the accuracy of the weights and consequently the survey estimates. We strongly advocate for the simultaneous performance of design weighting and sampling to prevent the loss of crucial information. Furthermore, this example underscores the critical importance of transparent sampling documentation for each sample in a (panel) survey, including frame and population sizes as well as a detailed description of every sampling step. A further explanation for the inability to fully replicate the findings from our synthetic data set when using actual panel data might be that the intersections of the different sampling frames, and thus the products of inclusion probabilities of the different recruitments in particular, have only a small impact on the estimates based on panel data with refreshments.

Despite encountering challenges in generating weights for application to the GESIS Panel data, the analysis of the synthetic data set demonstrates the necessity of employing multiple-frame weighting when integrating a refreshment sample into an ongoing panel. This study employed a multiple-frame approach

to recruiting respondents at three distinct points in time. Consequently, the variability observed in the design weights throughout the study did not reach a level requiring interventions such as trimming to reduce their variance. It is anticipated that a multiple-frame approach involving additional recruitments may substantially elevate the variability of the computed weights, leading to a corresponding increase in the variance of the design weights. Therefore, applying the multiple-frame approach to encompass all future waves will inevitably entail combining this procedure with a trimming approach to effectively mitigate the variability of the design weights.

The primary focus of this paper was on accurately calculating design weights in panel surveys with refreshment samples with the aim of yielding unbiased estimates in the absence of nonresponse and attrition. Consequently, we did not delve into the implementation of attrition and calibration weights. However, given that attrition is a primary driver for refreshing the panel population, it is essential to further examine the question of the optimal method for combining multiple overlapping frames and integrating attrition weights. Moreover, the multiple-frame approach discussed here aims to accurately compute inclusion probabilities used in a Horvitz-Thompson estimator, where the weights are inherently the inverse of the inclusion probabilities. Thus, the challenge lies in identifying an appropriate model specification to estimate attrition propensity rather than in the combination of the different frames.

References

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field Methods, 27*(4), 391–408. https://doi.org/10.1177/1525822X15574494

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. https://doi.org/10.1177/0894439317697949

Brick, J. M., Dipko, S., Presser, S., Tucker, C., & Yuan, Y. (2005). Estimation issues in dual frame sample of cell and landline numbers. In JSM Proceedings, Survey Research Methods Section (pp. 2791–2798). American Statistical Association. http://www.asasrms.org/Proceedings/y2005/files/JSM2005-000236.pdf

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2021). *Fortschreibung des Bevölkerungsstandes 2018, On-Site-Zugang*. https://doi.org/10.21242/12411.2018.00.00.1.1.0

Gabler, S., Häder, S., Lehnhoff, I. & Mardian, E. (2012). Weighting for unequal inclusion probabilities and nonresponse in dual frame telephone surveys. In S. Häder, M. Häder, & M. Kühne (Eds.), *Telephone surveys in Europe* (pp. 147–167). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-25411-6_11

GESIS – Leibniz Institute for the Social Sciences. (2021). *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften – ALLBUS sensitive Regionaldaten*. GESIS Data Archive. https://doi.org/10.4232/1.13767

Hartley, H. O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association, 19*(6), 203–206.

Heckel, C., & Wiese, K. (2011). Sampling frames for telephone surveys in Europe. In S. Häder, M. Häder, & M. Kühne (Eds.), *Telephone surveys in Europe* (pp. 103–119). Springer. https://doi.org/10.1007/978-3-642-25411-6_9

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society. Series A (General), 149*(1), 65–82. https://doi.org/10.2307/2981886

Lohr, S. L. (2007). Recent developments in multiple frame surveys. In *JSM Proceedings, Survey Research Methods Section* (pp. 3257–3264). American Statistical Association. http://www.asasrms.org/Proceedings/y2007/Files/JSM2007-000580.pdf

Lohr, S. L. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Statistics Canada – Survey Methodology, 37*(2), 197–213. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11608-eng.pdf?st=pwEXRPYa

Lohr, S., & Rao, J. N. K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association, 101*(475), 1019–1030. https://doi.org/10.1198/016214506000000195

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Statistics Canada – Survey Methodology, 33*, 151–157. https://www150.statcan.gc.ca/n1/pub/12-001-x/2007002/article/10492-eng.pdf

Sand, M. (2018). *Gewichtungsverfahren in Dual-Frame-Telefonerhebungen bei Device-Specific Nonresponse* (Doctoral dissertation; GESIS-Schriftenreihe No. 20). GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.21241/ssoar.60293

Sand, M., & Gabler, S. (2018). Gewichtung von (Dual-Frame -) Telefonstichproben. In Häder, S., Häder, M., & Schmich, P. (Eds.), *Telefonumfragen in Deutschland. Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute* (pp. 405–424). Springer VS. https://doi.org/10.1007/978-3-658-23950-3_13

Scherpenzeel, A. (2011). Data collection in a probability-based Internet panel: How the LISS panel was built and how it can be used. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 109*(1), 56–61. https://doi.org/10.1177/0759106310387713

Singh, A. C., & Mecatti, F. (2011). Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *Journal of Official Statistics, 27*, 633–650. https://www.proquest.com/scholarly-journals/generalized-multiplicity-adjusted-horvitz/docview/2821376567/se-2

Skinner, C. J., & Rao, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association, 91*(433), 349–356. https://doi.org/10.1080/01621459.1996.10476695

Wetzel, M., Schumann, N., & Schmiedeberg, C. (2021). *New weights for the pairfam anchor data* (pairfam Technical Paper No. 17). https://doi.org/10.5282/ubm/epub.91999

methods, data, analyses | Vol. 19(1), 2025, pp. 23-66 DOI: 10.12758/mda.2025.02

Probing in Cognitive Interviews Can Promote Acquiescence

Frederick G. Conrad[1], Rachel E. Davis[2], Carolyn Lau[3], Melissa Armendáriz[1], Stephanie Morales[1], Timothy P. Johnson[4] & Johnny Blair[5]

[1] *University of Michigan*
[2] *University of South Carolina*
[3] *Pew Research*
[4] *University of Illinois at Chicago*
[5] *Independent Consultant*

Abstract

Cognitive interviewing is widely used to pretest survey questionnaires and is considered a best practice (e.g., Willis, 2005, 2018; Beatty & Willis, 2007). However, the method has been controversial because, among other concerns, it requires interviewers to probe respondents for more detail or clarity about their experience answering draft survey questions which may lead them to report "problems" they have not actually experienced (e.g., Conrad & Blair, 2009). The present study investigates this possibility from the perspective of Acquiescent Response Style (ARS) – the tendency for survey respondents to select positive responses such as "yes" or "strongly agree," irrespective of the question's content (e.g., Baumgartner & Steenkamp, 2001). For example, respondents in a cognitive interview might affirm experiencing a problem mentioned in or implied by an interviewer's probe even if they have not actually experienced it. We embedded a probing experiment in a pretest of a health survey in which respondents participated in cognitive interviews that used either *directive* probes (n=41) or *non-directive* probes (n=26). Directive probes explicitly queried respondents about a specific, intentionally unlikely interpretation of each question in a draft questionnaire; non-directive probes were open-ended. Directive probe (DP) respondents affirmed the interpretation queried in the probes over five times more often than respondents in the non-directive probe (NP) group volunteered these interpretations. This pattern was reversed for interpretations of the questions that were volunteered, i.e., about which DP respondents were not asked: NP respondents volunteered alternative interpretations over four times more than DP respondents. These effects were particularly pronounced for respondents with lower levels of education and who were younger. The findings suggest that directive probing in cognitive interviewing can promote responding that is reminiscent of ARS – an *affirmation bias* – and likely harmful for the quality of evidence produced in cognitive interviews.

Keywords: cognitive interviews, directive probes, acquiescent response style, affirmation bias, acquiescence, verbal probes, satisficing

Acknowledgements

The research reported here was supported, in part, by National Cancer Institute Grant # 1 R01 CA172283-01A1. We thank Francesca Terzoli and Torie Harmon for coding transcribed interviews. The article is dedicated to the memory of our colleague and co-author, Johnny Blair, whose commitment to improving how research is done inspired us.

Direct correspondence to

Frederick G. Conrad, University of Michigan, Ann Arbor, United States
E-mail: fconrad@umich.edu

Cognitive interviewing is a widely used technique for pretesting questionnaires and considered essential for the creation of high-quality survey data (e.g., Miller, 2011; Ridolfo et al., 2020; Willis, 2005; Willis, 2015; Willis, 2018; Wright et al., 2021). There are many versions of cognitive interviewing, but the general approach is to conduct in-depth, semi-exploratory interviews about a questionnaire that is under development, during which respondents are asked to answer each question in a way that makes their thinking explicit, usually by thinking aloud and/or responding to verbal probes such as a request to paraphrase the question. The respondents' think-aloud protocols and responses to probes are examined for evidence that the respondent has encountered problems[1] answering the survey question – problems that would likely add measurement error to the resulting survey data if the questionnaire were to be used in production interviewing without further revision. For example, the respondent's verbal reports may make it evident that answering a question requires recalling events that are hard to distinguish from other similar events or that the respondent has misinterpreted the question relative to what the author intended. By rewording questions to resolve the problems uncovered in this way, researchers can minimize the chances that these same problems will introduce measurement error to the responses once production interviewing has begun, especially if the process is iterative.

The think-aloud procedure requires respondents to verbalize what is going through their minds while answering, but think-aloud protocols may not, by themselves, be interpretable by the cognitive interviewer, or whoever analyzes them (e.g., Ericsson & Simon, 1992). Thus, it is common for the cognitive interviewer to ask additional, often unscripted questions – what Willis (2005) calls "verbal probing" – to gain clarity about what the respondent may have meant. Several authors (e.g., Beatty, 2004; Conrad & Blair, 2009; Priede et al., 2014; Willis, 2005) have distinguished between different types of cognitive interviewing probes. The distinction most relevant to the current study is between probes that specifically refer to a problem that at least some respondents are anticipated to experience, e.g., "Did you think the question was asking about your expenses for prescription medicine?" and probes that do not refer explicitly to a possible problem, e.g., "What kind of expenses do you think that question was asking about?". The concern is that asking respondents whether they have experienced a particular problem, irrespective of evidence to that effect, may promote false

1 We use the term "problem" to cover both a respondent not understanding a question, i.e., they are confused by what they have been asked, and misunderstanding a question, i.e., the respondent interprets the question differently than its authors intended. In the former, they likely recognize their confusion and that they might be unable to respond, i.e., that they have encountered a problem. In the latter, they are likely unaware of the misalignment and thus unlikely to consider it a problem. For consistency with other research on cognitive interviewing which describes the evidence that a question is not functioning as intended as "problems," we use that term here.

reports of the problem (Conrad & Blair, 2009), which can lead to unnecessary changes to the question, wasting resources and potentially introducing new problems through the revision process.

Acquiescent Response Style. Respondents in production surveys – not necessarily pretests of questionnaires – sometimes endorse positive answer categories such as "Strongly Agree" irrespective of item content (e.g., Baumgartner & Steenkamp, 2001; Krosnick, 1991). This tendency, known as Acquiescent Response Style (ARS), is typically (although not always) observed when questions include bipolar scales (for example, a scale from "Strongly Agree" to "Strongly Disagree") used in questions about opinions and other subjective phenomena, primarily in interviewer-administered questionnaires (see Davis et al., 2024). There are several reasons why ARS may arise. First, the respondent may feel that selecting positive answers is more polite than selecting negative answers and may believe that being polite can facilitate the interaction with the interviewer. Second, in some cases, respondents may wish to avoid the effort required to carefully think through their answers and may want to do this without being conspicuous; selecting a response option because it is positive but without regard to what it means may satisfy both goals. Reducing the effort in this way can be seen as an instance of the more general tendency for some respondents to take mental shortcuts, referred to as *survey satisficing* (e.g., Krosnick, 1991; Roberts, *et al*., 2019), which also includes phenomena such as non-differentiation, rounded numerical answers, and primacy effects.

ARS is well known to be more common among Latinos than non-Latino whites (e.g., Aday et al., 1980; Liu et al., 2017; Ross & Mirowsky, 1984), possibly reflecting cultural factors such as *simpatía*, a Latino cultural value that promotes being pleasant, agreeable, likable, non-confrontational, and respectful in interpersonal interactions (e.g., Triandis et al., 1984; Davis et al., 2011). If so, this would be consistent with the politeness explanation for the phenomenon. We note that the prevalence of ARS has been found to differ between Latino subgroups (Davis, et al., 2019), presumably reflecting the considerable cultural, economic, and political diversity within the Latino population in the US (e.g., Zong, 2022).

ARS is also more common among respondents with less education (Liu et al., 2019; McClendon, 1991; Meisenberg & Williams, 2008; Messick & Frederiksen, 1958). Educational attainment is sometimes used as a proxy for "cognitive ability" or "cognitive sophistication" in studies of satisficing (e.g., Krosnick & Alwin, 1986); if that relationship extends to ARS (often considered a type of satisficing), it would be consistent with the effort reduction explanation as the response task is likely experienced as more difficult by those with lower levels of ability.

Also a possible indication of effort reduction, older respondents tend to show higher levels of ARS (e.g., Liu et al., 2019), potentially reflecting reduced aptitude due to cognitive aging and, thus, an impetus to simplify their task.

Whatever the origin of ARS, it is almost certainly a type of measurement error (Baumgartner & Steenkamp, 2001; Billiet & Davidov, 2008; Billiet & McClendon, 2000; Cheung & Rensvold, 2000; Hoffman et al., 2013; Weijters et al., 2008; Winkler et al., 1982). If respondents endorse statements they do not actually agree with or agree with less strongly than their response would indicate, this can distort survey estimates.

This article explores the possibility that an ARS-like process may be at play in cognitive interviews for pretesting questionnaires, contributing measurement error to the conclusions, much as ARS can distort the estimates and conclusions based on the data collected in production interviews. Because cognitive interviewers often probe respondents for more detail about their thinking than may be evident in their spontaneous verbalizations, respondents may affirm specific problems queried by probes much as they endorse positive response options irrespective of their actual opinions producing acquiescent survey responses.

Current Study

The current study investigates whether cognitive interview probes can lead respondents to agree with an interpretation of a question embodied in a probe, irrespective of whether they actually hold this interpretation. More specifically, the study asks whether and how often respondents in cognitive interviews agree with an interpretation mentioned in a probe even if that interpretation is implausible and unlikely to be arrived at spontaneously. Respondents were randomly assigned to one of two types of cognitive interview, either those in which interviewers administered *directive probes,* the Directive Probe (DP) group, or cognitive interviews in which the interviewer administered *non-directive probes,* the Non-directive Probe (NP) group. In the DP group, the questionnaire contained scripted probes, which asked the respondent if they interpreted the question in a specific – and unlikely – way; in the NP group, the scripted probes were open ended, asking respondents how they interpreted the question (see Table 1).

The critical aspect of directive probes as we define them here is that they ask the respondent to confirm or deny having experienced a specific problem, i.e., they are in effect Yes/No questions. In the current study the directive probes were scripted ahead of time, however cognitive interviewers as experts may – and, in our experience do – sometimes spontaneously ask the respondent to confirm that they have experienced a problem. We designed the question interpretations about which directive probes were administered to be highly implausible so that it would be unlikely for respondents to come up with these interpretations left to their own devices.

In contrast, the probes in the NP group did not mention any specific interpretations. For example, in Table 1 the NP probe asks, "who were you picturing doing this judging" (and if the respondent did not answer, they were provided an exhaustive set of options). This contrasts with the directive probe, which explicitly asks the respondent if they interpreted the judging to have been done by "strangers," chosen because in the authors' judgment respondents would be more likely to think of family members or other acquaintances than strangers. Thus, for a NP respondent to report the "strangers" interpretation, they would have to have volunteered, as opposed to being asked about, an unlikely interpretation. These NP probes are more like the kind of probes that are described in the cognitive interviewing literature. For example, Beatty and Willis (2007) propose a taxonomy of probes that are non-directive in that they do not refer to specific problems.

Further, requiring respondents in the NP group to volunteer their own interpretation provided evidence on whether the interpretations in the DP group were in fact unlikely. That is, if respondents in the NP group were to rarely volunteer an interpretation that was directly queried in the DP group, this would help confirm that the interpretation we designed into a directive probe was not the modal interpretation and so its affirmation by DP participants would raise concern about the veracity of their affirmation.

We note that the tasks that respondents in the DP and NP groups were asked to carry out were not identical. In the DP group, the task relied primarily on recognition[2], while the NP task relied primarily on recall: a DP respondent must determine whether the probed interpretation matches what is currently in mind while an NP respondent must articulate how they interpreted the question without any potential cues from the probe.

Hypotheses

H1a: Respondents in the DP group will be more likely to affirm the interpretation mentioned in the directive probes than will be NP respondents to volunteer that interpretaton. Thus, for a NP respondent to report interpreting the question in the same way described in the corresponding directive probe, the respondent would have had to reach the same unlikely interpretation explored in the directive probe, without it being mentioned.

2 The non-directive probes provided to interviewers included a version (in parentheses to indicate they were optional) that listed a relatively exhaustive set of response options. This was done so that if NP respondents were silent after being probed, they still had a chance to report their interpretation by selecting one of these options. When this option was exercised by the interviewer, it converted the respondent's task from primarily recall to primarily recognition. It still differed from the DP task in that the options were substantive, not "yes" or "no."

H1b: Respondents in the NP group will be more likely to provide an interpretation that is not mentioned in the corresponding directive probe than will DP respondents. This assumes that, without being asked about the interpretation described in the directive probes, NP respondents are unlikely to spontaneously arrive at that interpretation. If this is the case, then they will report an alternative interpretation.

Table 1 Examples of directive and non-directive probes. The material in parentheses after the non-directive probe was intended to be read only if the respondent was having trouble answering the probe.

Question	How important is it to you that people are judged by their own personal actions, and not by the actions of other people in their families? Is this not important, a little important, important, or extremely important?
Directive Probe	When you answered this question, did you think primarily about the judgments of strangers?
Non-directive Probe	When you answered this question, who were you picturing doing the judging? (Were you primarily thinking about close family and friends, acquaintances, strangers, some combination of these types of people, or someone else?)

If affirming the interpretation proposed in a directive probe is analogous to ARS, then this behavior should be more likely for the same subgroups who exhibit more ARS.

The evidence that Latinos tend to display high levels of acquiescence (e.g., Aday et al., 1980; Liu et al., 2017; Ross & Mirowsky, 1984) which varies between Latino subgroups (Davis et al., 2019), leads to the following hypotheses:

H2a: Latino respondents – in particular Cuban Americans, Puerto Ricans, and Mexican Americans – will affirm the interpretation proposed in directive probes more than will non-Latino White respondents.

H2b: Latino respondents – in particular Cuban Americans, Puerto Ricans, and Mexican Americans – will be less likely to offer an alternative response than will non-Latino Whites.

Further, the evidence that lower levels of education are associated with higher levels of ARS (Liu et al., 2019; McClendon, 1991; Messick & Frederiksen, 1958) leads to Hypotheses 3a and 3b:

H3a: Respondents with less formal education will affirm the interpretation proposed in directive probes more than will more highly educated respondents.

H3b: Respondents with less formal education will offer an alternative response less often than will more highly educated respondents.

Lastly, the evidence that older respondents tend to engage in ARS at higher rates than younger respondents (Lechner & Rammstedt, 2015; Lechner et al., 2019; Liu et al., 2019; Meisenberg & Williams, 2008) leads to the following hypotheses:

H4a: Older respondents will affirm the interpretation proposed in directive probes more than will younger respondents.

H4b: Older respondents will offer an alternative interpretation less often than will younger respondents.

Method

The experiment was embedded within a cognitive interview pretest of a questionnaire about Latino health conducted in the United States. Two versions of the questionnaire were tested over three rounds of cognitive interviews in English (n=86) and Spanish (n=37). A total of 45 closed-form questions covering a wide variety of topics including family relations, female gender roles, male gender roles, personal beliefs, assorted opinions, and cultural values, along with directive or non-directive probes, were included in versions of the cognitive interview guide for each probe condition. The questions for which probes were developed appear in Appendix A along with the probes for the corresponding probe group. In total, 123 in-person cognitive interviews were conducted at the University of Illinois Chicago Survey Research Laboratory. Respondents were randomly assigned to either the DP or NP group[3]. The resources available to the current project allowed us to transcribe and analyze 67 audio-recorded cognitive interviews, randomly selected from the larger pool.

Respondents. We recruited the respondents by placing ads on Craigslist and in local Spanish-language newspapers as well as by posting ads on listservs and flyers in neighborhoods with a high proportion of Latino residents. In addition, respondents recruited other potential respondents by word of mouth. Finally, staff called landline telephone numbers from samples believed to overrepresent Latino households, although in the end very few respondents were recruited from this sample source. Potential respondents completed a telephone screener in which they were asked about their ethnicity, gender, preferred language, and education, among other attributes. Additionally, respondents' level of acculturation (High Bicultural, Moderate Bicultural, Strong Latino, Latino-Leaning Bicultural, Anglo-Leaning Bicultural, Unclassified) was measured using the ARSMA-II (Bowman, 2005; Cuellar, I., et al., 1995), which was adapted slightly for use with an expanded set of Latino heritage groups. Only about 3% of respondents

3 Although each recruited sample member was randomly assigned to be interviewed following the DP or NP protocol, disproportionately more DP interviews were ultimately completed.

scored "Strong Latino," i.e., high Latino, low Anglo; other high Latino respondents were also high or moderate Anglo, thus placing them in the high or moderate Bicultural categories. Taken together this pattern of acculturation suggests that this sample was relatively acculturated (see Table 2).

Table 2 Respondent acculturation levels

Acculturation	Frequency	Percent
Strong Latino (high Latino, low Anglo)	2	2.99
Latino leaning bicultural (high Latino, moderate Anglo)	23	34.33
Moderate bicultural (moderate Latino, moderate Anglo)	14	20.90
High bicultural (high Latino, high Anglo)	6	8.96
Anglo leaning bicultural (moderate Latino, high Anglo)	12	17.91
Unclassified	10	14.93
Total	67	100

The cognitive interviews were conducted in the participant's preferred language: if they preferred "Only Spanish" or spoke "Spanish better than English" the interview was conducted in Spanish; if they reported preferring "Only English" or speaking "English better than Spanish" the interview was conducted in English; and if they answered "Both Spanish and English" the interview language was chosen at random. Eligible participants self-identified as a member of one of four groups: Mexican American, Puerto Rican, Cuban American, or non-Latino White (see Table B1).

Participants were each paid $50 upon completing the interview.

Of the 67 cognitive interviews analyzed in the current study, 41 (61%) were DP, and 26 (39%) were NP interviews[4]. Respondents were randomly assigned to type of cognitive interview so that ethnicity, gender, and interview language were roughly balanced between the two probe conditions (see description of these variables in "Analytic Approach" below). The distributions were indistinguishable between the two probe conditions: ethnicity, $\chi^2(1) = 0.39$, ns; gender, $\chi^2(1) = 0.27$, ns; interview language, $\chi^2(1) = 0.33$, ns.; and (while not deliberately balanced) education $\chi^2(1) = 0.63$, ns. The distributions of ethnicity, gender, interview language, and education across the two probe conditions in the 67 cognitive interviews are presented in Appendix B (Tables B1, B2, B3, and B4, respectively).

Both a directive and nondirective probe were constructed by the study team for each of the 45 questions in the draft questionnaire (see Appendix A).

4 The sample of 67 cognitive interviews was selected blind to the type of interview; thus the greater number of DP than NP interviews in the sample reflects the greater proportions of DP interviews in the total pool.

The directive probes were asked about a specific and, in the authors' view, unlikely interpretation. The authors' judgment about the likelihood of respondents interpreting the question in this way was confirmed by the low rate at which these interpretations were volunteered by the NP respondents (see Table 5). Many of the directive probes were designed to make the target interpretation unlikely by asking if it was the respondent's only interpretation, e.g.,

Q: How much do you believe that women should be comfortable voicing their opinions to men?

DP: When you answered this question, did you think only about when women have opinions about things that affect their families?

If respondents had interpreted the question to include the probed interpretation and others, they were free to indicate this, and this would have been coded as "Partial Affirm." This was rare for DP respondents (Table 5), suggesting that when they endorsed the probed interpretation, they were reporting it as their only interpretation.

The non-directive probes included an alternative version (in parentheses) that offered the respondent an exhaustive set of interpretations including the DP interpretation and, usually, an open option, e.g., "or something else?" In other words, while still non-directive the alternative version of each non-directive probe provides options from which respondents could choose. Interviewers were instructed to administer this version of the probe when NP respondents seemed unable to answer.

Interviewers. Seven interviewers conducted the cognitive interviews (see Table 3 for interviewer characteristics). Because the current study was embedded within an actual pretest for a production survey, the assignment of respondents to interviewers was driven largely by deadlines of the parent project and thus which interviewers were available when a respondent was recruited. No records were maintained of which interviewers conducted which cognitive interviews (as is the norm, in our experience, in actual pretests). As a result, we do not know whether an interviewer conducted DP, NP or both kinds of cognitive interviews.

The interviewers were trained in general interviewing techniques, cognitive interviewing techniques, and study-specific content. In the training sessions, the purpose of the main study (not the current cognitive interviewing study) was presented to the interviewers. Interviewers were reminded what cognitive interviewing is and the differences between cognitive interviewing and standardized field interviewing. Further, interviewers were instructed to read a "How To" guide (Willis, 1999) that covered cognitive interviewing techniques including probing, background theory, examples, and detection of problems. The interviewers were instructed to read each question as worded, to ask respondents to answer each question and report on their thinking, and after the respondent had both answered the question and reported on their thinking to read the scripted probe (whether directive or non-directive).

Table 3 Cognitive interviewer characteristics

Gender	6 female, 1 male
Native Language	5 bilingual, 2 English only
Latino ethnicity	5 Latina, 2 white
Profession	4 professional survey interviewers, 2 survey research supervisors, 1 PhD level social scientist

Questions. Twelve of the draft survey questions were designed to be asked only of males in the production interview and 15 were designed to be asked only of females in the production interviews: thus, in the cognitive interviews male respondents were asked 30 questions with probes and female respondents were asked 33 questions with probes. This created 1930 responses to the probes (1842 of which were codable) from the 67 cognitive interviews, all of which were analyzed in the current study.

Behavior Coding. All 67 cognitive interview audio recordings were transcribed and, if the interview was conducted in Spanish (n=28), first translated into English. Each transcribed interview was then coded by two independent judges for respondent and interviewer behaviors. A coding scheme, consisting of five behavior codes, was developed to classify respondents' answers to the probes based on the initial coding of 16 transcripts by one of the authors. The codes are presented in Table 4. The coding task was divided among two pairs of coders; one pair coded one set of arbitrarily selected interviews, and the other pair coded the remainder. Inter-rater reliability (κ) was computed across 66 interviews (one interview was used as a training case). The κ score was 0.83 indicating "strong" (McHugh, 2012) or "nearly perfect" (Everitt & Haye, 1992) agreement between the coders. After the κ score was calculated, the coders reconciled any differences in the codes they assigned so that one set of codes was available for analysis.

Note that NP respondents could not explicitly reject the DP interpretation: because their task was to report how they interpreted the question, not whether their understanding of the question matched one proposed by the researcher, they could offer an alternative interpretation, thus implicitly rejecting the DP interpretation; DP respondents could explicitly reject the probed interpretation by responding "no" when directly probed. Further, this means that DP respondents could, at their discretion, also offer an alternative; if they did volunteer an alternative interpretation, this was coded as "Provide Alternative" not "Reject." Thus, for NP respondents, offering an alternative and affirming or partially affirming the DP interpretation exhausted the possible responses to the (non-

Table 4 Behavior Codes

Code	Probe Condition	Description	Example (hypothetical)
Reject	DP (only)	R answered "no" to the directive probe	I: When you answered this question, did you think primarily about the judgments of strangers? R: No
Affirm	DP	R affirmed interpretation questioned in the directive probe and did not provide additional interpretation	I: When you answered this question, did you think primarily about the judgments of strangers? R: Yes
	NP	R volunteered the interpretation about which DP Rs were explicitly asked	I: When you answered this question, who were you picturing doing the judging? R: I was imagining strangers
Partially affirm	DP	R affirmed interpretation questioned in the directive probe and volunteered additional interpretation(s)	I: When you answered this question, did you think primarily about the judgments of strangers? R: Yes, and I also thought of family members
	NP	R volunteered interpretation about which DP Rs were explicitly asked and volunteered additional interpretation(s)	I: When you answered this question, who were you picturing doing the judging? R: I was imagining strangers and I also thought of family members
Provide Alternative	DP	R rejected the interpretation in the directive probe and provided alternative interpretation	I: When you answered this question, did you think primarily about the judgments of strangers? R: No, I thought of family members
	NP	R provided interpretation about which DP Rs were not explicitly asked	I: When you answered this question, who were you picturing doing the judging? R: I thought of family members.
Not codable	Observed only in NP	R's answer to the probe did not make sense or was not responsive.	I: When you answered this question, did you think ONLY about times when a woman is in physical danger? R: Generally.

directive) probe but this was not the case for DP respondents because they could also explicitly reject it.

The coding categories further distinguished between affirming (or volunteering for NP respondents) the DP interpretation without volunteering another interpretation (Affirm) and affirming the DP interpretation as well as volunteering at least one other interpretation (Partially Affirm). We made this distinction because we believed that affirming both the DP interpretation and at least one other suggested a weaker endorsement of the former than if it were the only interpretation endorsed.

Analytic Approach

Dependent Variables

We modeled the number of times each respondent affirmed or partially affirmed the probe for each question (i.e., probe), treating the composite variable simply as "Affirmations" (see Statistical Analysis) The other dependent variable that we modeled was "Provide Alternative," which was the number of times (questions) that each respondent offered an interpretation other than the one queried in the directive probe (excluding the alternative interpretations mentioned in partial affirms[5]). Both dependent variables were calculated at the respondent level.

Independent Variables

Probe group was a binary variable, DP or NP. Ethnicity was treated as a categorical variable in the models: Mexican American, Puerto Rican, Cuban American, and non-Latino Whites[6] (distributions of respondents across ethnic subgroups within each probe group appear in Table B1, Appendix B). Respondents' ages, which ranged from 20 to 67 years, were recoded into a categorical variable (20 – 34; 35 – 54; 55 years and older). Educational attainment was represented as a binary variable that distinguished between those with less than a bachelor's degree and those with a bachelor's degree or higher.

Statistical Analysis

To test our hypotheses, we first fit Poisson regression models to the data for affirming the DP interpretation and for providing an alternative interpretation – both of which are counts – using the glm function in R. While appropriate for

5 Because responses to the probes could only be assigned to one category, and affirming the directive probe is assumed to be measurement error, we prioritized its detection by treating partial affirmations as affirming the directive probe.

6 Note that the number of non-Latino Whites was small, n=7 in the DP group and n=3 in the NP group.

this type of count data, Poisson regression requires that the variance equal the mean (Ver Hoef & Boveng 2007), which was not the case for either dependent variable, i.e., the data were overdispersed, by the odTest function in the pscl package in R. Thus, we fit negative binomial regression models, using the glm.nb function that is part of the MASS package in R. Negative binomial regressions are appropriate for count data and can be fit despite overdispersion. The models tested the effect of probe group, ethnicity, age, and education on the number of affirmations or alternative interpretations provided by each respondent. One set of models (3 and 4) includes the interaction of probe group and education.

Results

Table 5 displays the percent of probes which each respondent, on average, affirmed, partially affirmed, implicitly rejected by offering of an alternative and, for the DP group, explicitly rejected. It was possible that DP respondents could have overwhelmingly rejected the interpretations queried in the probes (i.e., by responding "no") given the relative implausibility of these interpretations. This was not the case. DP respondents, on average, rejected fewer than half (41.6%) of the probed interpretations. Instead, they affirmed or partially affirmed the probed interpretation roughly as often (45.2%) as they rejected it. If this level of affirmation reflects the rate of actual question interpretation, as opposed to ARS-like behavior, then NP respondents should have volunteered (implicitly affirmed) those interpretations in similar proportions. This was also not the case. NP respondents, on average, affirmed (i.e., volunteered) only 9.6% of the DP interpretations. Instead, they volunteered an alternative interpretation for 74.1% of the (non-directive) probes, that is, when asked how they interpreted the question, on average three-quarters of their interpretations differed from the NP interpretation. In contrast, DP respondents offered an alternative to only 13.3% of the (directive) probes. In other words, DP respondents affirmed the interpretation queried in the probes four times more than NP respondents volunteered these interpretations, and NP respondents volunteered alternative interpretations of the questions more than five times as often as did DP respondents.

Note that DP respondents exclusively affirmed the probe about six times as often as they partially affirmed it, i.e., also affirmed at least one other interpretation (38.6% vs. 6.6%). In contrast, NP respondents volunteered (affirmed) only the DP interpretation *less* than they partially affirmed that interpretation, i.e., endorsed the DP interpretation and also offered at least one other interpretation (9.6% vs. 13.2%). Thus, it appears that directive probes greatly restricted how DP respon-

dents understood the questions or, perhaps more plausibly, their willingness to diverge from the probed interpretation.

Table 5 Mean percent of probes that each respondent affirmed, partially affirmed, implicitly rejected by providing an alternative and, for DP respondents, explicitly rejected. (Standard deviation in parentheses.)

Probe Group	Response to Probe			
	Affirm	Partially Affirm	Provide Alternative	Reject
Directive	38.6 (0.54)	6.6 (0.24)	13.3 (0.35)	41.6 (0.55)
Nondirective	9.6 (0.20)	13.2 (0.29)	74.1 (0.59)	

Note: 3.1% of responses to nondirective probes were uncodable.

Overall, the patterns in Table 5 are consistent with both H1a and H1b. We test H1a and H1b more directly in Models 1 and 2 (Table 6). In Model 1, the greater frequency of Affirmations (pooled Affirms and Partial Affirms) for the DP compared to the NP respondents is highly significant, confirming Hypothesis 1a, and serving as a check on the probe manipulation: the DP interpretations were rarely volunteered by NP respondents, i.e., respondents who were free to report how they understood the questions without being offered an interpretation by the researchers. Similarly, the greater frequency with which NP respondents offered an alternative to the probed interpretation than did DP respondents is highly significant in Model 2, confirming H1b.

To the extent that affirming an unlikely interpretation resembles ARS, it would follow that Latino subgroups might exhibit more affirmation than non-Latino Whites (H2a). Our data do not support this hypothesis. Mexican Americans and Puerto Ricans in the DP group affirmed the probe no more often than non-Latino Whites, and Cuban Americans affirmed the probe significantly *less* often than non-Latino Whites (Model 1), in a reversal of what we predicted and what would be expected based on the ARS literature in which Latinos generally exhibit more ARS than non-Latino Whites (e.g., Aday et al., 1980; Liu et al., 2017; Ross & Mirowsky, 1984). Regarding H2b, we proposed that Latinos would be less likely to offer an alternative response than would non-Latino Whites. There was

no evidence in support of this prediction as shown in Model 2: the differences between the ethnic groups and non-Latino Whites were not significant.

Table 6 Negative binomial regression results for affirming the probe (Model 1) and offering an alternative interpretation (Model 2)

Variables	Model 1 (Affirming the Probe)			Model 2 (Offering an Alternative Interpretation)		
	B	*SE*	*p-value*	*B*	*SE*	*p-value*
Intercept	1.0189	0.2176	**<.0001**	3.0685	0.2131	**<.0001**
Directive	1.2111	0.1481	**<.0001**	-2.0189	0.1319	**<.0001**
Non-Directive (ref.)	.	.	.	.	.	.
Mexican/American	0.1252	0.2106	0.5523	-0.0722	0.2231	0.7460
Puerto Rican	-0.3467	0.2188	0.1130	0.2875	0.2192	0.1896
Cuban/American	-0.9268	0.2239	**<.0001**	0.0009	0.2149	0.9963
non-Latino White (ref.)	.	.	.	.	.	.
Age, years						
20-34 (ref.)	.	.	.	.	.	.
35-54	-0.2908	0.1441	**0.0436**	0.0960	0.1458	0.5104
≥ ≥ 55	0.0929	0.1666	0.5769	0.3875	0.1823	**0.0335**
Less than a bachelor's	0.6226	0.1599	**<.0001**	-0.2138	0.1423	0.1329
Bachelor's degree or higher (ref.)	.	.	.	.	.	.

More consistent with the ARS literature (e.g., Lechner et al., 2019; Meisenberg & Williams, 2008), education was a strong predictor of affirming the probe. Those without a bachelor's degree were significantly more likely to affirm the probe than those with a bachelor's degree or higher (Model 1), a finding that supports H3a. Given the strength of the education effect, we asked whether it was equally strong for both probe groups, i.e., was the tendency to affirm the probe in the DP group stronger for those with lower levels of education as the ARS literature would predict, without being similarly moderated by education in the NP group where the task was less likely to trigger ARS? Thus, we tested the interaction between probe group and educational attainment in Models 3 and 4. As shown in

Bachelor's degree or higher (ref.)	.	.	.	.	.	.
Directive * Less than a bachelor's degree	0.9976	0.3038	**0.001**	-0.1218	0.2739	0.6566

Table 7, the interaction is significant in Model 3: DP respondents with less than a bachelor's degree *were* significantly more likely to affirm the probe than those with more education but education made little difference in the NP group.

Table 7 Negative Binomial regression (including an interaction term) results for affirming the probe (Model 3) and offering an alternative interpretation (Model 4)

Variables	Model 3 (Affirming the Probe)			Model 4 (Offering an Alternative Interpretation)		
	B	*SE*	*p-value*	*B*	*SE*	*p-value*
Intercept	1.5682	0.2553	**<0.0001**	3.0242	0.2302	**<0.0001**
Directive Probe	0.5345	0.2435	**0.0281**	-1.9421	0.2132	**<0.0001**
Non-Directive (ref.)	.	.	.	.	.	.
Mexican/American	-0.0078	0.1989	0.9687	-0.0462	0.2275	0.839
Puerto Rican	-0.4362	0.2064	**0.0346**	0.3059	0.2200	0.1644
Cuban/American	-0.9506	0.2131	**<0.0001**	0.0101	0.2136	0.9624
non-Latino White (ref.)	.	.	.	.	.	.
Age, years						
20-34 (ref.)	.	.	.	.	.	.
35-54	-0.3078	0.1347	**0.0224**	0.0987	0.1451	0.4962
≥ 55	0.0692	0.2611	0.7718	0.3904	0.1816	**0.0316**
Less than a bachelor's degree	-0.7570	0.1599	**<.0001**	-0.1742	0.1678	0.2990

Figure 1 displays the average percentage of affirmations (affirms + partial affirms) for each respondent by probe group and education level. DP respondents with a bachelor's degree or higher affirmed the probe 31% of the time compared to those with less than a bachelor's degree who affirmed the probe 53% of the time. Differences are small and in the opposite direction in the NP group: those with a bachelor's degree volunteered the DP interpretation 26% of the time and those with less than a bachelor's degree volunteered that interpretation about as often, 21% of the time. Thus, those with less formal education are driving the increased number of affirms in the DP interviews[7].

7 When the significant education x probe type interaction is included in Model 3, the main effect of education is significant but reversed relative to its direction in Model 1. We attribute this to education moderating the main effect of probe type (the interaction of these two variables is significant) so that when the interaction is included in the model, the residual main effect of education is what "remains" after the interaction is removed, making it largely uninterpretable.

Education did not affect the frequency of offering an alternative response, leading us to reject H3b (Model 2) and the interaction between education level and probe group was not significant for offering an alternative response (Model 4). No other interactions were significant and so none are included in the models.

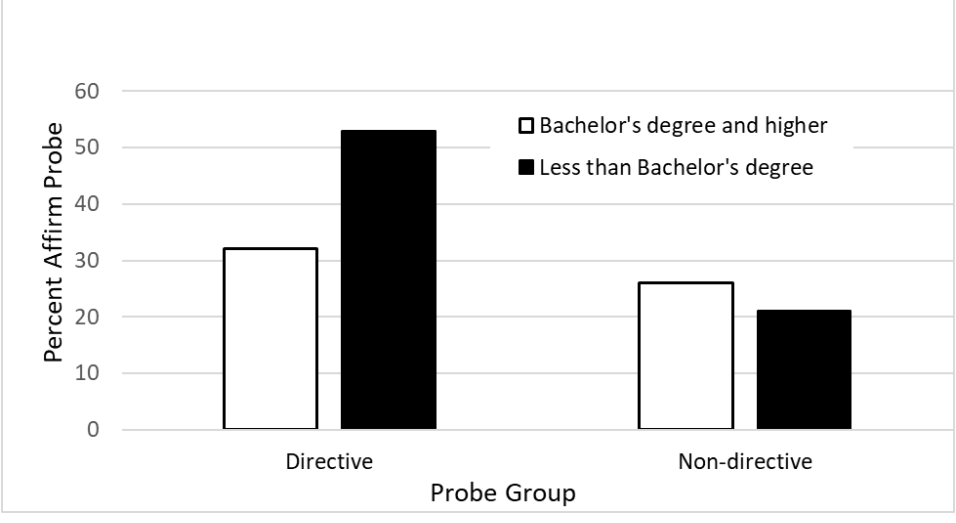

Figure 1 Percent of Affirmations (Affirm + Partial Affirm) for Probe by Education group.

Respondent age has been shown to increase ARS (e.g., Liu et al., 2019). Thus, we tested H4a (older respondents will affirm the probe more than younger respondents) in Models 1 and 3 and H4b (older respondents will offer an alternative less often than younger respondents) in Models 2 and 4. In all the models, the reference age was 20 - 34 years, the youngest group. Respondents between 35 and 54 years of age were *less* likely to affirm the probe than those under 34 years of age, reversing the typical finding in the ARS literature and contradicting H4a, significantly in Model 3 and marginally so in Model 1, while the differences between respondents 55 years and older and those 20 - 34 years of age were not statistically significant. Further, Models 2 and 4 show that those who are 55 years of age and older were significantly *more* likely to offer an alternative response in comparison to younger respondents, a reversal of the H4b prediction.

Finally, the reversal of Hypothesis 2a (more frequent affirmation of the probe among Latino subgroups) observed among Cuban Americans in Model 1 who affirmed the probe significantly less than did non-Latino Whites, is also observed among Puerto Ricans in Model 3. There is, as in Model 2, no support

for Hypothesis 2b (more frequently offered alternative responses) in Model 4; none of the Latino subgroups offered alternatives at different rates than did non-Latino Whites.

Discussion

Respondents in the DP group were substantially more likely to affirm they had interpreted a set of survey questions in the rare and unintended ways queried in directive probes than were NP respondents to volunteer those interpretations without being explicitly asked about them. Similarly, NP respondents volunteered alternative interpretations substantially more often than did DP respondents. Taken together, these findings provide strong support for our original intuition that cognitive interview results are vulnerable to error when respondents are directly asked if they experienced a particular problem, especially compared to those for whom the problem was not explicitly presented.

It is possible that these differences could be related to differences in reporting tasks: for the DP group the task had the character of a *recognition* task while respondents in the NP group were, in effect, asked to *recall* their interpretation (or when the parenthesized alternative non-directive probe was administered, to choose from a set of alternatives). Recall is generally more prone to error than recognition (e.g., Anderson, 2020; Tulving & Thompson, 1973), so it is possible that NP respondents interpreted the questions in much the same way as DP respondents but simply forgot their interpretation. This seems unlikely because the interval between the question's delivery to respondents and when they were probed for their interpretation was very brief, presumably too brief for much forgetting to have occurred.

It is also possible that DP respondents' affirmations accurately reflected their interpretations, i.e., that they did in fact interpret the questions in the improbable ways queried in the directive probes. This, too, is unlikely given how rarely NP respondents volunteered the same interpretations. Moreover, the far greater frequency with which NP respondents volunteered alternative interpretations should have been mirrored by DP respondents, keeping in mind that respondents were randomly assigned to one probe group or the other. That this was not the case raises the question of why DP respondents might have affirmed understanding questions in a way that may not have been entirely accurate.

We have suggested that the patterns of results are due to ARS-like processes. To the extent that ARS is a type of survey satisficing (e.g., Krosnick, 1991; Roberts, et al., 2019), i.e., respondents simplifying the task, especially if their ability is limited, or reducing effort when their motivation is low, affirming the probe may perform much the same function. Respondents' education level serves as a proxy for cognitive ability in the survey satisficing literature (e.g., Krosnick &

Alwin, 1987). In the current study, DP respondents with lower levels of education (less than a Bachelor's degree) were more likely to affirm the probe than those with more education, consistent with the satisficing view of ARS in the literature.

That the youngest respondents affirmed the probe more than the those in their middle years and offered an alternative less than respondents older than 55 years of age, reverses the general finding in the ARS literature and could argue against the ARS analogy we propose here. But it is consistent with greater survey satisficing by younger than older respondents (e.g., Anduiza & Galais, 2017; Liu et al., 2017; Zhang & Conrad, 2014), presumably reflecting younger respondents' reduced motivation.

To the extent that ARS is about getting along with conversational partners and lubricating the interaction involved in answering survey questions, we did not find evidence that affirming the probe served this purpose. Although the greater frequency of ARS among Latino than non-Latino White respondents has been attributed to *simpatía* – the cultural norm that promotes being pleasant, agreeable, likable, non-confrontational, and respectful in interpersonal interactions – two of the three Latino subgroups in the current study, Cuban Americans and, in one model, Puerto Ricans, affirmed the probe *less* than did non-Latino Whites, reversing the predicted effect; the remaining subgroup, Mexican Americans exhibited no more affirmation of the DP interpretation and volunteered alternative interpretations no less often than did non-Latino Whites. This result could reflect variation in cultural traditions between different groups of Latinos. Alternatively, it could be due to the relatively assimilated character of the Latino participants: most were at least moderately bicultural, with only two being classified as Strong Latino (Table 2). A less assimilated Latino sample might well have affirmed directly probed interpretations more than did the Latino respondents in the current study.

Related to ethnicity, the non-Latino Whites in the current study, having been recruited from sample sources believed to overrepresent Latinos, could have been more similar culturally to those who identified as Latino, e.g., they might have been highly assimilated Latinos who identify as White, than might non-Latino Whites from a more general sample source. But at least for our sample, the version of ARS we observed seems less related to *simpatía*, and more related to reducing effort when ability and motivation are lower.

Whatever the exact mechanism, directive probing appears to lead to an *affirmation bias*. It seems far easier to affirm a problem proposed by an interviewer than to generate a description of a different problem, potentially leading to false alarms (Conrad & Blair, 2009). This can certainly jeopardize the quality of information provided by cognitive interviews in which directive probes are administered, as well as the quality of survey data elicited after the questionnaire is revised – based on reported problems that include high levels of false alarms – and then administered in production research.

A clear practical implication of these findings is that interviewers should avoid directly probing specific problems in cognitive interviews. In some types of qualitative research, interviewers are authorized to confirm their understanding of the interview data with participants (e.g., Olson, 2016; Tracy, 2010), and as "detectives" (Willis 1994) cognitive interviewers may wish to unambiguously confirm their hypotheses about potential problems by directly asking respondents. However, the current results seriously question the wisdom of this approach, at least as the primary method of exploring respondents' interpretations in cognitive interviews. If DP respondents were willing to agree with the unlikely interpretations suggested by the interviewers, directive probes might be affirmed even more often if the problems they mention are more plausible. This is not to suggest that directive probing is always ill-advised. For example, to clarify whether they have correctly understood something the respondent reported or implied, interviewers might directly ask respondents to confirm their understanding of the respondent's interpretation (what Conrad & Blair, 2009, called a "conditional probe") but in the absence of evidence that the respondent has interpreted the question in a particular way, the results of directive probing seem likely to mislead designers revising a questionnaire.

Instead, probing "around" the problems (interpretations) whose presence the interviewer wishes to confirm without mentioning them explicitly can corroborate their existence without introducing affirmation bias. The probes that were administered in the NP group requested open responses in most cases and so did not imply to respondents that a particular problem was under investigation. Similarly, the many example probes provided by Willis (2005) and Beatty and Willis (2007) also have an open, non-directive character. To be clear, the current results do not bear on the merits of scripted versus improvised probing: whether probes are planned or developed on the fly, it is possible to understand how respondents have understood a question without asking them to affirm they have interpreted the question in a specific, problematic way.

Next steps. We noted that respondents from some Latino subgroups affirmed directive probes less often than did non-Latino Whites, a reversal of the pattern predicted by the ARS literature, and we suggested that this pattern might have been due to the predominant bicultural orientations of the Latino participants. A future study might investigate whether less acculturated respondents, i.e., those for whom *simpatía* is presumably more prominent, might exhibit higher levels of affirming the probed interpretation in the current study. Similarly, a comparison group of non-Latino Whites who are recruited from general sample sources rather than sources in which Latino representation is expected to be high, could sharpen the comparison. If Latino subgroups and especially the least assimilated members of those subgroups exhibit more affirmation of probed interpretations than non-Latino Whites, it could begin to suggest that ARS – at least the version of it that seems to reflect a desire to avoid controversy and negativity

– may be more prevalent in cognitive interviews than in those conducted for the current study.

Related to this, the sample of 67 cognitive interviews analyzed in the current study was large compared to typical pretests but not large enough for us to have full confidence that the differences in affirmations between Latino subgroups (particularly between Cuban Americans) and non-Latino Whites would hold up with larger samples (see Blair & Conrad, 2011). Increasing the number (and typicality) of non-Latino Whites in a future study could reveal effects of subgroup membership, which were not detected in the current study due to insufficient power.

A follow-up study would not only collect data from a larger number of respondents in each subgroup but would (1) recruit a larger number of cognitive interviewers and, (2) randomly assign them to conduct either DP or NP interviews. This would make it possible to explore interviewer clustering of affirmations, alternative interpretations, and in DP interviews, rejections of the DP interpretation. It is possible that because of differences in how individual interviewers administer the probes or even deliver the draft questions, different interviewers might elicit different patterns of responses to the probes, analogous to interviewer effects in standardized, production interviews (e.g., Fowler, Lewis, & Magione, 1992; Groves & Magilavey, 1986; West & Blom 2017).

Davis et al. (2019) report no evidence that interviewers' characteristics affected ARS in production interviews, possibly suggesting that the way they conduct cognitive interviews is not related to the kind of ARS-like behavior observed in the current study. Nonetheless, it would be worth testing if cognitive interviewers differ from each other in whether and how often they administer directive probes. If interviewer effects of this type are small this would bolster the current findings and suggest that it is possible for cognitive interviewers to consistently explore question understanding without probing specific interpretations.

Conclusion

There is little doubt that revising survey questionnaires based on pretests is a low-cost way to help assure that the data collected in production research are as high quality as possible. The current study adds complexity to this view by providing evidence that sometimes pretest results are themselves subject to measurement error, in this case an affirmation bias that is triggered by interviewers probing specific misinterpretations of questions in cognitive interviews. Taking steps to reduce this source of error by, for example, training cognitive interviewers to avoid directive probing, seems likely to succeed. But it suggests that researchers may need to be more discriminating in how they interpret the results of cognitive interviews before revising questions based on those results

and, more generally, to examine what other types of error in cognitive interview data may be attributable to how those interviews are conducted.

References

Aday, L. A., Chiu, G. Y., & Andersen, R. (1980). Methodological issues in health care surveys of the Spanish heritage population. *American Journal of Public Health, 70*(4), 367-374. https://doi.org/10.2105/AJPH.70.4.367

Anderson, J. R. (2020). *Cognitive psychology and its implications* (9th ed.). Worth Publishers.

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research, 29*(3), 497-519. https://doi.org/10.1093/ijpor/edw007

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143 - 156. https://doi.org/10.1509/jmkr.38.2.143.18840

Beatty, P.C. (2004). The dynamics of cognitive interviewing. In Presser, S. Rothgeb, J.M., Couper, M.P., Lessler, J.Y., Martin, E., Martin, J., and Singer, E. (Eds.). *Methods for testing and evaluating survey questionnaires* (pp. 45-66). John Wiley and Sons.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), 287-311. https://doi.org/10.1093/poq/nfm006

Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research, 36*(4), 542-562. https://doi.org/10.1177/0049124107313901

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*(4), 608-628. https://doi.org/10.1207/S15328007SEM0704_5

Bauman, S. (2005). The reliability and validity of the brief acculturation rating scale for Mexican Americans–II for children and adolescents. *Hispanic Journal of Behavioral Sciences, 27*(4), 426-441. https://doi.org/10.1177/0739986305281423

Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly, 75*(4), 636-658. https://doi.org/10.1093/poq/nfr035

Cuellar, I., Arnold, B., & Maldonado, R. (1995). Acculturation rating scale for Mexican Americans-II: A revision of the original ARSMA scale. *Hispanic Journal of Behavioral Sciences, 17*(3), 275-304. https://doi.org/10.1177/07399863950173001

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*(2), 187-212. https://doi.org/10.1177/0022022100031002003

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly, 73*(1), 32-55. https://doi.org/10.1093/poq/nfp013

Davis, R. E., Conrad, F. G., Dong, S., Mesa, A., Lee, S., & Johnson, T. P. (2024). An ounce of prevention: using conversational interviewing and avoiding agreement response scales to prevent acquiescence. *Quality & Quantity, 58*(1), 471-495. https://doi.org/10.1007/s11135-023-01650-7

Davis, R. E., Johnson, T. P., Lee, S., & Werner, C. (2019). Why do Latino survey respondents acquiesce? Respondent and interviewer characteristics as determinants of cultural patterns of acquiescence among Latino survey respondents. *Cross-Cultural Research, 53*(1), 87-115. https://doi.org/10.1177/1069397118774504

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Revised Edition)*. MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Everitt, B. S., & Hay, D. F. (1992). *Talking about statistics: A psychologist's guide to data analysis*. Halsted Press.

Groves, R. M., & Magilavy, L. J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly, 50*(2), 251-266. https://doi.org/10.1086/268979

Hoffmann, S., Mai, R., & Cristescu, A. (2013). Do culture-dependent response styles distort substantial relationships? *International Business Review, 22*(5), 814-827. https://doi.org/10.1016/j.ibusrev.2013.01.008

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201-219. https://doi.org/10.1086/269029

Lechner, C. M., & Rammstedt, B. (2015). Cognitive ability, acquiescence, and the structure of personality in a sample of older adults. *Psychological Assessment, 27*(4), 1301–1311. https://doi.org/10.1037/pas0000151

Lechner, C. M., Partsch, M. V., Danner, D., & Rammstedt, B. (2019). Individual, situational, and cultural correlates of acquiescent responding: Towards a unified conceptual framework. *British Journal of Mathematical and Statistical Psychology, 72*(3), 426-446. https://doi.org/10.1111/bmsp.12164

Liu, M., Conrad, F. G., & Lee, S. (2017). Comparing acquiescent and extreme response styles in face-to-face and web surveys. *Quality & Quantity, 51*, 941-958. https://doi.org/10.1177/10731911211042932

Liu, M., Suzer-Gurtekin, Z., Keusch, F., & Lee, S. (2018). Response styles in cross-cultural surveys. In Johnson, T. P., Pennell, B. E., Stoop, I. A., & Dorer, B. (Eds.). *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (477-499). *John Wiley & Sons.* https://doi.org/10.1002/9781118884997

Mangione, T. W., Fowler, F. J., & Louis, T. A. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics, 8*, 293-293. https://doi.org/10.1093/jssam/smz031

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*(7), 1539-1550. https://doi.org/10.1016/j.paid.2008.01.010

Messick, S., & Frederiksen, N. (1958). Ability, acquiescence, and "authoritarianism". *Psychological Reports, 4*(3), 687-697. https://doi.org/10.2466/pr0.1958.4.3.687

McClendon, M. J. (1991a). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research, 20(1), 60-103.* https://doi.org/10.1177/0049124191020001003

McClendon, M. J. (1991b). Acquiescence: Tests of cognitive limitations and question ambiguity hypotheses. *Journal of Official Statistics, 7(2), 153-166.* https://doi.org/10.1037/met0000631

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22(3), 276–282.* https://doi.org/10.11613/BM.2012.031

Miller, K. (2011). Cognitive interviewing. In Maddans, J., Miller, K., Maitland, A. & Willis, G. (eds.), *Question evaluation methods (pp. 51-75). Hoboken, NJ: John Wiley & Sons.*

Olson, K. (2016). *Essentials of qualitative interviewing*. Routledge. https://doi.org/10.4324/9781315429212

Priede, C., Jokinen, A., Ruuskanen, E., & Farrall, S. (2014). Which probes are most useful when undertaking cognitive interviews? *International Journal of Social Research Methodology, 17(5), 559-568.* https//doi.org/10.1080/13645579.2013.799795

Ridolfo, H., Ott, K., Beach, J., & McCarthy, J. S. (2020). Pre-testing establishment surveys: Moving beyond the lab. *Survey Practice, 13*(1), 11810. https://doi.org/10.29115/SP-2020-0003

Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly, 83*(3), 598-626. https://doi.org/10.1093/poq/nfz035

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior,* 189-197. https://doi.org/10.2307/2136668

Tracy, S. J. (2010). Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry, 16*(10), 837-851. https://doi.org/10.1177/1077800410383121

Triandis, H. C., Marin, G., Lisansky, J., & Betancourt, H. (1984). Simpatía as a cultural script of Hispanics. *Journal of Personality and Social Psychology, 47*(6), 1363. https://doi.org/10.1037/0022-3514.47.6.1363

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352-373. https://doi.org/10.1037/h0020071

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology, 88*(11), 2766-2772. https://doi.org/10.1890/07-0043.1

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*(3), 409-422. https://doi.org/10.1007/s11747-007-0077-6

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology, 5(2), 175-211.* https://doi.org/10.1093/jssam/smw024

Willis, G. B. (1999). Reducing Survey Error through Research on the Cognitive and Decision Processes in Surveys. https://www.hkr.se/contentassets/9ed7b1b3997e4bf4baa8d4eceed5cd87/gordonwillis.pdf.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design. Sage Publications.* https://doi.org/10.4135/9781412983655

Willis, G. B. (2015). The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly, 79*(S1), 359-395. https://doi.org/10.1093/poq/nfu092

Willis, G. B. (2018). Cognitive interviewing in survey design: State of the science and future directions. In D. Vannette and J. Krosnick (Eds.), *The Palgrave handbook of survey research*, 103-107. https://doi.org/10.1007/978-3-319-54395-6_14

Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology, 67*(5), 555-561. https://doi.org/10.1037/0021-9010.67.5.555

Wright, J., Moghaddam, N., & Dawson, D. L. (2021). Cognitive interviewing in patient-reported outcome measures: A systematic review of methodological processes. *Qualitative Psychology, 8*(1), 2–29. https://doi.org/10.1037/qup0000145

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods* 8(2),127-135.

Zong, J. (2022). A mosaic, not a monolith: A profile of the US Latino population, 2000 – 2022. Latino Policy & Politics Institute, UCLA. https://latino.ucla.edu/research/latino-population-2000-2020/

Appendix A

Questions with Probes.

Note that Non-Directive probes included in parentheses a version with an relatively exhaustive set of options so that respondents could choose one of several options if they were otherwise silent.

(How important is it to you that) teenage children be encouraged to develop their independence? (Is this…)

1 ❐ not important,
2 ❐ a little important,
3 ❐ important, or
4 ❐ extremely important?

[PROBE: DIRECTIVE] When you answered this question, did the word "encourage" make you think about providing rewards to children, such as sweets or present, when they show independence?
[PROBE: NON-DIRECTIVE] When you answered this question, in what ways were you thinking that teenage children would be encouraged to develop their independence? (Did the word "encourage" make you think about encouraging independence by providing opportunities, praise, rewards such as sweets or presents, some combination of these things, or something else?)

(How important is it to you that) people are judged by their own personal actions, and not by the actions of other people in their families? (Is this…)

1 ❐ not important,
2 ❐ a little important,
3 ❐ important, or
4 ❐ extremely important?

[PROBE: DIRECTIVE] When you answered this question, did you think primarily about the judgments of strangers?
[PROBE: NON-DIRECTIVE] When you answered this question, who were you picturing doing the judging? (Were you primarily thinking about close family and friends, acquaintances, strangers, some combination of these types of people, or someone else?)?

[TO BE ADMINISTERED TO FEMALE PARTICIPANTS ONLY]

How much do you believe that women are more responsible than men for taking care of the emotional needs of their families? Would you say you…

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, were you thinking primarily of those situations in which someone in the family is upset?
[PROBE: NON-DIRECTIVE] When you answered this question, what did "taking care of emotional needs" mean to you? (Did "taking care of emotional needs" make you think primarily about situations in which someone in the family was upset, making sure people are happy on a day-to-day basis, both of these situations, or something else?)

How much do you believe that a woman should think of others' needs before her own? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about what a woman thinks, regardless of how she acts?
[PROBE: NON-DIRECTIVE] When you answered this question, were you thinking only about how women think, or also how they act?

How much do you believe that women should be comfortable voicing their opinions to men? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think only about when women have opinions about things that affect their families?
[PROBE: NON-DIRECTIVE] When you answered this question, what kinds of opinions were you thinking about? (Were you thinking about women's opinions about things that affect their families or their opinions in general?)

How much do you believe that a woman has to be strong to be successful

in life? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think that "strong" means that a woman insists on things being done her way?
[PROBE: NON-DIRECTIVE] What does the word "strong" mean to you in this question? (Did the word "strong" make you think about a woman insisting on getting her way, being physically strong, not showing fear, some combination of these things, or something else?)

How much do you believe that important decisions should be made by the man of the household? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, were you thinking that the man would only make important decisions after consulting his wife?
[PROBE: NON-DIRECTIVE] When you answered this question, were you thinking that the man would make important decisions all on his own, after talking with his wife, after talking with other family members, some combination of these actions, or something else?

How much do you believe that a woman should be free to make up her own mind? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think only about major decisions that a woman makes in her lifetime such as whether or not to get married or work outside the home?
[PROBE: NON-DIRECTIVE] What kinds of things were you thinking that a woman would make up her mind about when you answered this question? (Were you thinking about major decisions that a woman makes in her lifetime such as whether or not to get married or work outside the home, more minor decisions such as what clothes to buy, her opinions about things in general such as what

she thinks about climate change, people, or movies, or some combination of these types of things?)

How much do you believe that a woman should never show fear? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about times when a woman is in physical danger?
[PROBE: NON-DIRECTIVE] What kinds of situations were you thinking about when you answered this question? (Were you thinking only about situations in which a woman simply feels nervous or uncomfortable such as when she is talking in front of a group of people, only situations in which she is in physical danger, both of these types of situations, or something else?)

How much do you believe that a woman should obey her husband's wishes? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about things that a husband feels very strongly about?
[PROBE: NON-DIRECTIVE] What kinds of "wishes" were you thinking about when you answered this question? (Were you thinking about a husband's wishes about small things, things that he feels very strongly about, both of these types of wishes, or something else?)

How much do you believe that although the man may not know it, the decisions are really made by the woman of the house? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about small decisions?
[PROBE: NON-DIRECTIVE] What kinds of decisions were you thinking about when you answered this question? (Were you thinking about only small decisions, only big decisions, or decisions in general?)

How much do you believe that women hold the most power within their households? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about a woman's power over the children in the household?
[PROBE: NON-DIRECTIVE] When you answered this question, which people were you thinking about when it comes to a woman's power? (Were you thinking mostly about how much power a woman has over her husband, her children, other people in the household, or some combination of these types of people?)

How much do you believe that a woman needs to be strong willed to gain the respect of others? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did the word "others" make you think ONLY about men?
[PROBE: NON-DIRECTIVE] Who did you think of as the "others" when you answered this question? (Were you thinking mostly about men, mostly about women, or both men and women?)

How much do you believe that women should be in charge of making their own decisions about their lives? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about minor decisions?

[PROBE: NON-DIRECTIVE] What kinds of decisions were you thinking about when you answered this question? (Were you thinking about minor decisions, only about major decisions, or all types of decisions?)

How much do you believe that it is the responsibility of the woman in the household to set a moral example for her family to follow? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think that "family" referred mainly to the woman's husband?
[PROBE: NON-DIRECTIVE] Which people in the family do you think this question is asking about? (Were you thinking mainly about the woman's husband, mainly about her children, someone else, or some combination of these people?

How much do you believe that child care should primarily be a woman's responsibility? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think it was asking who decides how children should be cared for, regardless of who actually provides the care?
[PROBE: NON-DIRECTIVE] When you answered this question, what did "child care" mean to you? (Did you think the question was asking who decides how children should be cared for, who actually provides the care, both making decisions and providing the care, or something else?)

How much do you believe that a woman should not let others tell her what to do? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think that it was mostly asking about strangers?

[PROBE: NON-DIRECTIVE] When you answered this question, who were the "others" that you think this question was asking about? (Did you think that this question was mostly asking about a woman's husband, her children, her parents, her co-workers, strangers, someone else, or some combination of people?)

[TO BE ADMINISTERED TO MALE PARTICIPANTS ONLY]

How much do you believe that a man should be affectionate with his children? Would you say you…

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you compare how affectionate men and women are with their children?
[PROBE: NON-DIRECTIVE] When you answered this question, were you thinking only about men or were you comparing men and women?

How much do you believe that a man should not let others tell him what to do? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think that it was mostly asking about co-workers?
[PROBE: NON-DIRECTIVE] When you answered this question, who were the "others" that you think this question was asking about? (Did you think that this question was mostly asking about a man's wife, his children, his parents, his co-workers, strangers, someone else, or some combination of people?)

How much do you believe that a man should never show fear? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about times when a man is in physical danger?
[PROBE: NON-DIRECTIVE] What kinds of situations were you thinking about when you answered this question? (Were you thinking only about situations in which a man simply feels nervous or uncomfortable such as when he is talking in front of a group of people, only situations in which he is in physical danger, both of these types of situations, or something else?)

How much do you believe that it is necessary for a man to fight when challenged? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about older men?
[PROBE: NON-DIRECTIVE] When you answered this question, were you mostly thinking about younger men, middle-aged men, older men, some combination of ages, or men of all ages?)

How much do you believe that it is important for women to look good? (Would you say you…)

1 ☐ don't believe that at all,
2 ☐ believe that a little,
3 ☐ somewhat believe that, or
4 ☐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think primarily about whether women look clean and proper, as though they were going to church?
[PROBE: NON-DIRECTIVE] When you heard this question, what did "look good" mean to you? (Were you thinking whether women dress and do their hair in a sexy way, have a sexy figure, look clean and proper as though they were going to church, some combination of these things, or something else?)

How much do you believe that men cannot be expected to be as honorable

as women? (Would you say you…)

1 ❒ don't believe that at all,
2 ❒ believe that a little,
3 ❒ somewhat believe that, or
4 ❒ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about whether men could be as faithful to their wives and girlfriends as women are to their male partners?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of actions or qualities were you thinking about? (Were you thinking only about whether men could be as faithful to their wives and girlfriends as women are to their male partners, whether men are as religious as women, whether men are as moral as women, some combination of these things, or something else?)

How much do you believe that a man should be in control of his wife? (Would you say you…)

1 ❒ don't believe that at all,
2 ❒ believe that a little,
3 ❒ somewhat believe that, or
4 ❒ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about times when the man and his wife are out in public?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of situations were you thinking about? (Were you thinking mostly about times when the man and his wife are in the privacy of their home, at the homes of family or friends, out in public, in some other type of situation, or some combination of situations?)

How much do you believe that if a woman is being insulted, a man should defend her? (Would you say you…)

1 ❒ don't believe that at all,
2 ❒ believe that a little,
3 ❒ somewhat believe that, or
4 ❒ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about times when a woman is insulted by another woman?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of situations were you thinking about? (Were you thinking mostly about times when

a woman is insulted by another man, times when a woman is being insulted by another woman, times when a woman is being insulted by someone else, or anytime a woman is being insulted?)

How much do you believe that men should be in charge of the finances in their households? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about major financial decisions?
[PROBE: NON-DIRECTIVE] What types of finances or financial decisions were you thinking about when you answered this question? (Were you thinking only about minor financial decisions, only about major financial decisions, both minor and major types of financial decisions, or something else?)

How much do you believe that a man obtains honor from treating other people with respect? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about how a man treats strangers?
[PROBE: NON-DIRECTIVE] What types of "other people" were you thinking about when you answered this question? (Were you thinking about how a man treats his wife, his children, close family and friends, acquaintances, strangers, someone else, or some combination of these types of people?

How much do you believe that men should not talk about their feelings? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about whether or not men should talk about their feelings with their wives?
[PROBE: NON-DIRECTIVE] When you answered this question, which people were you thinking about in terms of whom men should talk or not talk to about their feelings? (Were you thinking about whether or not men should talk about their feelings with their wives, their children, other close family or friends, acquaintances, strangers, someone else, or some combination of these types of people?)

How much do you believe that a man should respect a woman's opinion, regardless of her age? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about situations when the man is at work?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of settings were you thinking about? (Were you thinking mostly about private conversations between a man and his wife, social settings with friends and families, situations when the man is at work, something else, or some combination of these types of situations?)

How much do you believe that there are many things we have not discovered yet, so nobody can be absolutely certain that their beliefs are right? Would you say you…

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about the things that people learn through their personal experiences?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of discoveries were you thinking about? (Were you thinking mostly about the things that people learn through their personal experiences throughout their lives, mostly about discoveries made by scientists, both of these types of things, or something else?)

How much do you believe that it is good to be open-minded? (Would you say you…)

1 ❒ don't believe that at all,
2 ❒ believe that a little,
3 ❒ somewhat believe that, or
4 ❒ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about your willingness to have new experiences?
[PROBE: NON-DIRECTIVE] When you answered this question, what kinds of things were you thinking about being open-minded about? (Were you thinking mostly about your willingness to have new experiences, to meet new people, to accept new ideas, something else, or some combination of these types of things?)

How much do you believe that you are certain that your ideas about the central issues in life are correct? (Would you say you…)

1 ❒ don't believe that at all,
2 ❒ believe that a little,
3 ❒ somewhat believe that, or
4 ❒ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about having a happy family?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of "central issues in life" were you thinking about? (Were you thinking mostly about having a happy family, the way in which the world works, the meaning of life, the existence of God, whether human nature is essentially good or bad, something else, or some combination of these types of things?)

How much do you believe that it is better to risk saying too much than to risk being misunderstood? (Would you say you…)

1 ❒ don't believe that at all,
2 ❒ believe that a little,
3 ❒ somewhat believe that, or
4 ❒ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about times when you are talking with someone who you know well?

[PROBE: NON-DIRECTIVE] When you answered this question, were you thinking mostly about times when you are talking with strangers, acquaintances, people who know each other well, or some combination of these types of people?

How much do you believe that how something is said generally communicates more information than the words used to say it? (Would you say you…)

1 ❐ don't believe that at all,
2 ❐ believe that a little,
3 ❐ somewhat believe that, or
4 ❐ believe that very much?

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about how the gestures that people use when they talk?
[PROBE: NON-DIRECTIVE] What does "how something is said" mean to you in this question? (Were you thinking mostly about how the gestures that people use when they talk, about how a person emphasizes certain words, how a person uses facial expressions to communicate meaning, something else, or some combination of things?)

In general, white Americans treat Latinos with respect.

1 ❐ 1 — STRONGLY DISAGREE
2 ❐ 2
3 ❐ 3
4 ❐ 4
5 ❐ 5
6 ❐ 6
7 ❐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you think ONLY about whether or not white Americans are polite, such as saying "please" or "thank you"?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of situations were you thinking about? (Were you thinking ONLY about whether or not white Americans are polite such as saying "please" or "thank you, whether they support laws and policies that help Latinos such as immigration reform and access to education and health care, whether they hire Latinos for jobs, something else, or some combination of things?)

Gay marriage should be illegal.

1 ❐ 1 — STRONGLY DISAGREE

2 ❐ 2
3 ❐ 3
4 ❐ 4
5 ❐ 5
6 ❐ 6
7 ❐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you think about men marrying men as well as women marrying women?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of couples were you thinking about? (Were you thinking mostly about men marrying men, women marrying women, both, or something else?)

If I were to choose a snack from a store, I would probably choose something sweet.

1 ❐ 1 — STRONGLY DISAGREE
2 ❐ 2
3 ❐ 3
4 ❐ 4
5 ❐ 5
6 ❐ 6
7 ❐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you think that "something sweet" included fruits such as bananas or pineapple?
[PROBE: NON-DIRECTIVE] When you answered this question, what kind of snacks were you thinking about? (Were you thinking about fruits such as bananas or pineapple, sugary snacks such as candy bars, both of these types of snacks, or something else?)

I enjoy watching reality TV shows.

1 ❐ 1 — STRONGLY DISAGREE
2 ❐ 2
3 ❐ 3
4 ❐ 4
5 ❐ 5
6 ❐ 6
7 ❐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you include nature shows?

[PROBE: NON-DIRECTIVE] When you answered this question, what types of TV shows were you thinking about? (Were you including only reality TV shows, or did you also include news shows, nature shows, educational TV shows, something else, or some combination of shows?)

I enjoy cold weather.

1 ❐ 1 — STRONGLY DISAGREE
2 ❐ 2
3 ❐ 3
4 ❐ 4
5 ❐ 5
6 ❐ 6
7 ❐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you compare how much you enjoy cold weather versus hot weather?
[PROBE: NON-DIRECTIVE] What were you thinking about when you answered this question? (Were you thinking only about how much you enjoy hot weather or did it make you compare how much you enjoy cold weather versus hot weather?)

I always remain calm during a crisis.

1 ❐ 1 — STRONGLY DISAGREE
2 ❐ 2
3 ❐ 3
4 ❐ 4
5 ❐ 5
6 ❐ 6
7 ❐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] Did you include heated arguments between people as "crises" when you answered this question?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of crises were you thinking about? (Were you thinking that "crises" referred only to emergencies, or did you also include heated arguments between people or other types of situations?)

I enjoy listening to stories.

1 ☐ 1 — STRONGLY DISAGREE
2 ☐ 2
3 ☐ 3
4 ☐ 4
5 ☐ 5
6 ☐ 6
7 ☐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] Did you include listening to jokes when answering this question?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of stories were you thinking about? (Were you thinking only about stories that people tell, jokes, books on tape, some other type of stories, or some combination of these types of things?)

Gays should have the same marriage rights as straight men and women.

1 ☐ 1 — STRONGLY DISAGREE
2 ☐ 2
3 ☐ 3
4 ☐ 4
5 ☐ 5
6 ☐ 6
7 ☐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about the rights and privileges that come with being married, such as visiting a spouse in the hospital?
[PROBE: NON-DIRECTIVE] When you answered this question, what did "marriage rights" mean to you? (Were you thinking mostly about the right to get married, the rights and privileges that come with being married such as visiting a spouse in the hospital, both of these types of rights, or something else?)

When it comes to interacting with Latinos, most white Americans are

racist.

1 ☐ 1 — STRONGLY DISAGREE
2 ☐ 2
3 ☐ 3
4 ☐ 4
5 ☐ 5
6 ☐ 6
7 ☐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about how white Americans react to hearing the opinions of Latinos on the TV, radio, or internet?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of interactions were you thinking about? (Were you thinking mostly about when white Americans are talking face-to-face with Latinos, mostly about how white Americans react to hearing the opinions of Latinos on the TV, radio, or internet, something else, or some combination of types of interaction?)

I like to spend time outdoors every day.

1 ☐ 1 — STRONGLY DISAGREE
2 ☐ 2
3 ☐ 3
4 ☐ 4
5 ☐ 5
6 ☐ 6
7 ☐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about active outdoor activities, such as walking?
[PROBE: NON-DIRECTIVE] When you answered this question, what types of activities were you thinking about? (Were you thinking of doing any particular activities – such as walking or gardening – or just sitting and enjoying the outdoors?)

Being around other people gives me energy.

1 ☐ 1 — STRONGLY DISAGREE
2 ☐ 2
3 ☐ 3
4 ☐ 4
5 ☐ 5
6 ☐ 6
7 ☐ 7 — STRONGLY AGREE

[PROBE: DIRECTIVE] When you answered this question, did you think mostly about feeling emotionally energized?
[PROBE: NON-DIRECTIVE] When you answered this question, what type of energy were you thinking about? (Were you thinking mostly about physical energy, mental energy, spiritual energy, something else, or some combination of types of energy?)

Appendix B

Distribution of Demographic Characteristics in the Two Probe Groups

Table B1 Latino ethnicity

Probe Group		Latino Ethnicity				
		Non-Latino White	Mexican-American	Puerto-Rican	Cuban American	Total
Directive	n	7	12	12	10	41
	%	17.1	29.3	29.3	24.4	100
Non-Directive	n	3	8	8	7	26
	%	11.5	30.8	30.8	26.9	100
Total	n	10	20	20	17	67
	%	14.9	29.9	29.9	25.4	100

Table B2 Gender

Probe Group		Gender		
		Male	Female	Total
Directive	n	21	20	41
	%	51.2	48.8	100
Non-Directive	n	15	11	26
	%	57.7	42.3	100
Total	n	36	31	67
	%	53.7	46.3	100

Table B3 Interview language

Probe Group		Interview Language		
		English	Spanish	Total
Directive	n	25	16	41
	%	60.9	39.0	100
Non-Directive	n	14	12	26
	%	53.9	46.2	100
Total	n	39	28	67
	%	58.2	41.8	100

Table B4 Education

Probe Group		Education					
		Less than High School	High School graduate	Some College	Bachelor's Degree	Graduate Degree	Total
Directive	n	5	9	10	9	6	39
	%	12.8	23.0	25.6	23.1	15.4	100
Nondirective	n	5	6	6	6	3	26
	%	19.2	23.1	23.1	23.1	11.5	100
Total	n	10	15	16	15	9	65
	%	15.4	23.1	24.6	23.1	13.9	100

Mother Tongue or Non-Native Language? – The Influence of Language on Response Behavior in Surveys

Florian Heinritz

Leibniz Institute for Educational Trajectories &
Universität Hamburg

Abstract

Today, an increasing number of surveys offer respondents the choice of which language they want to answer the questionnaire. In later data analysis, however, the language in which the respondent answers the questions is often ignored, and no distinction is made regarding whether that language is the respondent's mother tongue. Several psychological theoretical considerations and empirical observations indicate that respondents' answering behaviors are influenced by whether the questions are presented in their mother tongue or a non-native language. Therefore, the extent to which these mechanisms and effects of language used are also applicable and relevant in social science studies remains unclear. Based on models of cognitive load, satisficing, and language-dependent memory, the influence of language nativeness on response behavior is explained from a theoretical point of view. The research question will be answered by analyzing the data from the refugee study ReGES (Refugees in the German Educational System). The results of the analyses show that there is a difference in response behavior depending on whether a question is answered in a mother tongue or a non-native language. The implications, both from a survey methodological point of view and for further research, will be discussed.

Keywords: non-native language effect; language; multilingual surveys; response behavior; refugees

Due to the increase in labor migration (International Labour Office [ILO], 2010, 2018) and the number of refugees (International Organization for Migration [IOM], 2019), multilingual interviews recently became more relevant. Since multilingual interviews have already been conducted commonly in many countries with multiple national languages, the methodological challenges of conducting identical questionnaires in different languages (e.g., Hunt & Bhopal, 2004; McKay et al., 1996; Pan et al., 2014) and other methodological aspects of multilingual surveys have already been investigated in detail (e.g., Blohm & Diehl, 2001; Dotinga et al., 2005; Schoua-Glusberg, 2004).

However, even if these methodological challenges are considered, there can be differences in answering questions depending on the language used (e.g. Peytcheva, 2018). Nevertheless, research regarding language differences has often focused on bilingual respondents. But the following three developments raise the relevance for shifting the focus: First, due to the increase in migrants, there is an increasing group of people whose mother tongue is not one of the national languages of a country. Second, the number of forced migration is increasing at the same time, which typically means that people cannot properly prepare their migration by, for example, learning the national language of the host country. And third, relatively new survey technologies, such as multilingual computer- or web-based questionnaires, make it possible that the number of languages offered in a survey from which the respondent can choose no longer depends on the interviewer's language skills, as is usually the case in interviews with interviewers.

The interaction of these three points—the greater diversity of different mother tongues within countries, the rising number of people without knowledge of the national languages and the technical possibility for respondents to select their preferred language for responding to questionnaires or individual questions from a range of languages—opens a new question: Does it make a difference whether respondents answer a question in their mother tongue rather than in a non-native language? Analyses of numerous psychological studies and experiments suggest that there is a difference between answering a question in a non-native language as opposed to the mother tongue (for a summary see Hadjichristidis et al., 2019). However, whether and how these effects are also relevant in surveys has rarely been analyzed (e.g., Kappelhof, 2017). Therefore, the aim of this paper is, on the one hand, to investigate whether differences in response

Notes

The project ReGES is funded by the German Federal Ministry of Education and Research under grant number FLUCHT03. However, the authors have sole responsibility for the content of this publication.

Direct correspondence to

Florian Heinritz, Leibniz Institute for Educational Trajectories & Universität Hamburg, E-Mail: florian.heinritz@lifbi.de

behavior depend on whether questions are posed and answered in a mother tongue instead of a non-native language. On the other hand, the influence of language nativeness will also be analyzed for other practical aspects relevant for surveys, such as the duration of an interview or the accuracy of statements.

For this purpose, data from a German refugee survey ReGES (Will et al., 2021) are used to analyze the extent to which the language used influences the length of the survey, the accuracy of the information provided, and the actual response. Analyzing the data from this refugee study enables an investigation of the impact of language on response behavior in actual surveys based on computer-assisted self-interviews (CASI) in eight languages and—since the refugees are all newly arrived immigrants—allows a clear distinction between mother tongue and non-native language. Conversely, the data is not based on an experimental study.

Having this in mind, theoretical models and explanations are presented, and hypotheses are formulated in a first step. Subsequently, I briefly describe the data and the operationalization of the subsequent analyses before the results of the multivariate analyses are presented. The results show that items are answered more quickly in a mother tongue than in a non-native language. Likewise, by looking at items on gender roles and religiosity, it turns out that items about social norms are answered more in accordance with the norms associated with the mother tongue if these items are answered in a mother tongue. Finally, the results are discussed in a concluding section.

Theoretical Background and Previous Research

The following theoretical considerations focus on why people answer questions in a survey differently in their mother tongue than in a non-native language. As mentioned in the introduction, this paper is about people who have one or more mother tongues and have later learned a non-native language.

From a survey methodological point of view, the influence of language on response behavior can be explained by the model of satisficing (Krosnick, 1991). In addition, I will focus on two psychological approaches to explain why response behavior may change depending on the language used to answer the question: cognitive load theory and language-dependent memory. As shown in Figure 1, I will apply these models to the four survey methodology-relevant steps of answering a question, which are the comprehension of the question, retrieval of relevant information, judgment, and response (Tourangeau, 1984).

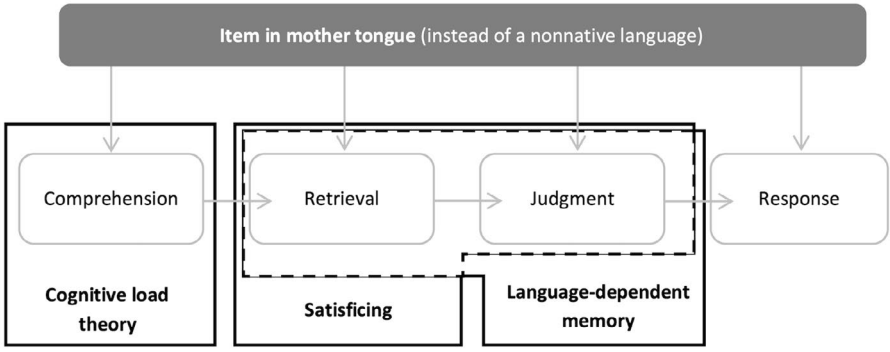

Figure 1 Overview of theories and their link to the respondent's task in the response process

While each theoretical model is used to derive a hypothesis in the following sections, it should be noted that the three theoretical models are not based on contradictory assumptions but rather complement each other.

Cognitive Load Theory

The initial task of the respondent in the process of answering a question is to comprehend the question or, more precisely, to understand the text (Tourangeau, 1984). Whereas it can be assumed that—despite different language skills (and depending on the survey mode used also despite different reading skills)—respondents have sufficient language skills to understand questions posed in their mother tongue, language skills of a respondent in a non-native language usually vary greatly.

Especially in the case of the group of refugees focused on in this paper, it can be assumed that the non-native language skills of the host country's language are lower than in the mother tongue, since refugees often have to leave the country unexpectedly and sometimes do not yet know in which country they will be placed. This makes texts in a non-native language comparatively harder to comprehend. Additionally, it is necessary to consider that items in surveys are even more difficult to understand because respondents usually cannot understand the meaning of the text from the context but have to understand each item individually (Calderón et al., 2006, pp. 50–51).

However, even if people understand a non-native language almost as well as their mother tongue, it is still the case that the cognitive load is higher for understanding a non-native language than a mother tongue (e.g., Hasegawa et al., 2002). Since cognitive load is defined as the amount of working memory

resources used to complete a mental activity such as comprehending a text (e.g., Paas & Van Merriënboer, 1994; Sweller, 1994) and these human memory resources are limited, a high cognitive load also means a higher cognitive effort (e.g., Paas et al., 2003).

As a result, comprehension of the questions is more challenging in a non-native language, either due to a lack of language proficiency or the higher cognitive load required to understand questions in a non-native language. The time required to answer items should vary depending on the language used. Therefore, respondents should need less time to respond to questions posed in their mother tongue than they would to questions posed in a non-native language:

H1: *Answering questions in a mother tongue takes less time than answering questions in a non-native language.*

Satisficing

As a model based on a combination of cognitive load theory and general rational choice theory (see Esser, 1990), satisficing can also be used to explain differences in response behavior based on language in the following two steps: retrieval and judgment (Tourangeau, 1984). According to rational choice theory (RCT), respondents evaluate in the answering process of each question how the highest possible subjective expected utility can be achieved by answering the question. However, since neither cost nor negative sanctions nor particularly high gains are expected in a scientific, voluntary survey, the subjective expected utility is evaluated as relatively low. Therefore, the use of cognitive load is often reduced to a minimum, meaning that respondents put less effort into answering the question when the task is difficult. This could, of course, be done by responding to the question in the mother tongue. However, this paper does not delve further into the reasons for answering a question in a non-native language when the mother tongue is available.

This minimal cognitive effort, which is also known as satisficing (Krosnick, 1991), can influence response behavior by reducing the cognitive load. One possible consequence of satisficing is reducing the cognitive load through heaping (Gideon et al., 2017). This means that the respondents try to minimize their cognitive effort by rounding open numerical answers instead of choosing the more cognitively demanding process of intensively retrieving the exact number and judging whether the number given is actually correct. Therefore, rounded answers are less accurate than unrounded answers (Battisin et al., 2003). A problematic consequence of this on data report and analysis is that rounded or estimated (and thus less accurate) answers given by the respondent can lead to a loss of validity. Therefore, it is also important to minimize the measurement error of heaping.

According to satisficing theory, the difficulty of the task is a significant factor influencing satisficing and, consequently, heaping (Krosnick, 1991). As previously stated in the considerations for Hypothesis 1, it can be assumed that answering a question in a non-native language increases the task difficulty and thus fosters satisficing. Therefore, the following relationship is predicted in Hypothesis 2:

H2: *Items presented in a mother tongue tend to be answered more accurately.*

Language-Dependent Memory

A second model that often serves as a possible explanation in this context is language-dependent memory (e.g., Marian & Neisser, 2000), which is applied here to the two steps of retrieval and judgment. This model relies on different circumstances of language learning. For example, the emotional context of the learning hypothesis states that a language is associated with emotions when that language is learned and used in an emotional context (Harris et al., 2006). It can therefore be assumed that the mother tongue, which is learned in childhood under the influence of many emotions, is much more strongly linked to emotions than a non-native language, which is usually learned in a (less emotional) educational context. The relationship of social norms and language is similar (see Nichols et al., 2016): Social norms are mainly internalized in the mother tongue and are thus more activated by the mother tongue than in a non-native language. Therefore, such experiences and learned norms are stored in long-term memory in the language in which the experiences were made or norms were acquired (e.g., Marian & Fausey, 2006; Marian & Kaushanskaya, 2004). These theoretical assumptions refer not only to differences between the mother tongue learned in childhood and a later learned non-native language but also, in the case of bilingual persons, to memories, norms or emotions associated with different languages (e.g., Danziger & Ward, 2010; Dewaele & Nakano, 2012; Marian & Kaushanskaya, 2004).

The theory of language-dependent memory implies that emotions and social norms play a much greater role in the mother tongue than in a non-native language, in which "cool-headed responses toward certain moral dilemmas" and "less condemnation of moral and social violations" can be expected (Hadjichristidis et al., 2019, p. 264). Therefore, depending on the emotions and social norms associated with a language, questions can be answered differently according to the language in which they are presented, and norms are less activated in a non-native language (Geipel et al., 2015).

For these reasons, it can be assumed that those questions, where norms or emotions linked to a language have to be retrieved and judged in the answering process, will be answered differently in a mother tongue or a non-native

language. This means that respondents would answer such questions in their mother tongue more emotionally and with the knowledge of social norms linked to their mother tongue. In a more formally learned non-native language, the associated social norms would be less activated, and emotionality would also decrease. It can be assumed that questions about social norms or emotions in a non-native language are less likely to be answered in accordance with the social norms and emotions incorporated in the mother tongue but rather answered more rationally and therefore more in conformity with the social desirability resulting from the interview context (e.g., the culture of the country where the interview takes place). Hypothesis 3 states therefore:

H3: *If items about social norms or emotions are presented in a mother tongue, the answers will be answered more in accordance with the norms associated with the mother tongue.*

Current State of Research

The question of how language influences response behavior has been addressed by many studies in different disciplines, each with a different focus. However, almost no study has analyzed the actual influence of the use of mother tongue or non-native language on response behavior in surveys. To obtain a sense of the empirical evidence supporting the hypotheses, some studies that have dealt with assumptions similar to the abovementioned hypotheses are briefly presented below.

In the United States, where a large proportion of the population is bilingual, studies have more frequently investigated the effect of language on response behavior in surveys (e.g., Diaz-Morales et al., 2006; Guarnaccia et al., 1989; Pérez, 2009; Welch, 1973). However, they have made no distinction between mother tongue and non-native language. Most of these studies show that the response behavior depends in different ways on whether the questionnaire was completed in English or another language.

As one of the few studies that focused on differences between mother tongue and non-native language, Harzing and Maznevski (2002) showed in an experiment with students that questions are answered in accordance with the cultural values linked to a language. This finding corresponds to other studies (e.g., Lee, 2001; Marin et al., 1983), although they did not distinguish between mother tongue or non-native language, and is in line with the theoretical assumptions of language-dependent memory theory. Similarly, the results from Kappelhof (2017), using data from a Dutch study on ethnic minorities, show that individuals answer items on family ties more traditionally in their mother tongue.

In addition, various psychological experiments have focused explicitly on the differences in response behaviors between respondents using their mother

tongue or a non-native language (for an overview, see Hadjichristidis et al., 2019). Some studies have shown that people perceive non-native languages less emotionally (e.g., Caldwell-Harris & Ayçiçeği-Dinn, 2009; Harris et al., 2003). Other studies have also shown that decisions in non-native languages were therefore made more rationally and less emotionally (e.g., Cipolletti et al., 2015; Costa et al., 2014; Geipel et al., 2015; Hadjichristidis et al., 2019; Hayakawa & Keysar, 2018; Shin & Kim, 2017).

In sum, these studies show that parts of the assumptions have already been empirically proven, so the theoretical explanations are a useful basis for the assumptions in the individual hypotheses. To test the hypotheses empirically, the data for the later analyses will be described in the following section.

Data and Methods

Unlike many recent studies on differences due to mother tongue and non-native language, the hypotheses are tested using a large-scale study. Specifically, data from the first wave from 2018 of the German refugee study ReGES "Refugees in the German Educational System" are used. Since a considerable number of children and young refugees came to Germany in the context of asylum immigration in the mid-2010s, this study focuses on the educational trajectories of young refugees by interviewing adolescents and parents, even though the sampling units were young refugees children (at least four years old but not attending school at that time) and refugees adolescents (between 14 to 16 years old) (Will et al., 2021). The target population was sampled via a complex, multi-stage sampling process from the German registration office across five federal states in Germany. For this purpose, a random sample of the nationalities of the most common refugee nationalities in Germany at that time was taken (for details, see Steinhauer et al., 2018). Respondents were contacted personally by interviewers after receiving an invitation letter, resulting in 5,711 completed interviews in the first wave (for further details regarding the sample over the waves, see Heinritz & Will, 2021 and von Maurice & Will, 2023).

To prevent panel conditioning, which may also influence response behavior, only the first wave of the study ReGES is considered. The advantage of this study is that the questionnaires of the first wave were offered in eight different languages: English, German, Arabic, Kurmanji, Pashto, Tigrinja, Farsi, and French. The languages—in addition to German as the original language of the study—were chosen based on the most common official languages of the respondents' countries of origin (Gentile et al., 2019), knowing that they are not always the mother tongues of the respondents. In all these languages, native speakers of the respective languages were employed as interviewers. In order to contact the respondent in the correct language, the nationality of the respondents was

used as an indicator for the language in which the interviewer should contact the person (e.g., respondents from Syria should be contacted by Arabic-speaking interviewers).

Sample

The main part of the interviews was a CASI. As it could be assumed that there were many illiterate people in the sample, the CASI was offered with audio files and the interviewers were also allowed to read the question aloud to the respondents. In order to minimize the possible influences and interactions of the interviewer[1], only those CASI interviews are analyzed in which the interviewers did not read out questions (n = 2,031). In addition, there were some cases (n = 84) where the respondents stated that they could not read, but neither used audio files nor asked the interviewer for help. These implausible cases are excluded as well as cases with implausible data on the mother tongues. This automatically excluded people with poorer reading skills in the analyses sample. Therefore, a total of 1,865 persons are considered in the following analyses. Furthermore, since it can be expected that social desirability differs depending on the country of origin or culture (Tourangeau & Yan, 2007, p. 860), the test of Hypothesis 3 includes only persons from Syria as the largest group of the sample with the same country of origin and therefore similar cultural background.

Table 1 provides a first description of the sample that served as the basis for the multivariate analyses. When looking at the sample in the second column in Table 1, which reports the mother tongues (multiple answer) of the respondents, it can be observed that the languages used correspond to the official languages of the countries of origin of the sample: Arabic (as the official language in Syria and Iraq) is the most common mother tongue, with 81.29% of respondents listing this language as one of their mother tongues, followed by Kurmanji (as an official language in Iraq and spoken in parts of northern Syria) and Farsi (as an official language in Iran), with 8.26%. Although Kurmanji was the mother tongue of many respondents, items views in this language cannot be included in the following analyses, as the complete translation of Kurmanji had to be revised during the fieldwork and quality problems remained due to the complexity of the language (see Gentile et al., 2019).

In ReGES, respondents were free to choose the language used for answering. On the one hand, the respondents could choose the interview language at the beginning; on the other hand, the language could be changed individually for

1 Interviewers can influence response behavior in many ways. These includes characteristics of the interviewer such as ethnicity, gender or age (e.g., Glantz and Michael, 2014; Groves et al., 2009; Loosveldt, 2008).

Table 1 Distribution and use of language in the sample in percent

	Mother tongues of respondents	Languages used for answering	Language matches in each language
Units	Respondents in analyses sample	Screens of analyses sample	Screens of analyses sample displayed in the language
Arabic	81.29	73.78	90.99
German	1.39	19.38	3.75
Farsi	8.26	6.03	98.94
English	7.56	0.71	24.86
Tigrinya	0.16	0.06	100
Pashto	0.70	0.08	100
Kurmanji	19.46	*	*
French	0.59	-	-
N	1,865	219,325	219,325

Source: ReGES data, own calculations, Wave 1.
* Language was not considered in the analysis sample.

each question.[2] The third column in Table 1 illustrates the proportion of screens displayed in each language. For example, 73.78% of the screens were last displayed in Arabic and 6.03% in Farsi, which roughly reflects the proportion of people with these languages as mother tongues. This distribution of mother tongues seems to correlate in most cases with the languages actually used in the survey (see the second column in Table 1). In fact, in 90.99% of the screens that were answered in Arabic, Arabic was also the mother tongue, and in Farsi, it was also the mother tongue in 98.94% of the cases (see the fourth column in Table 1). The greatest difference between the number of native speakers and number of users of this language in the interviews can be seen for German. Although only 1.39% of the respondents stated that German was their mother tongue, 19.38% of the screens were answered in German. Therefore, it is not surprising that only 3.75% of the items answered in German were answered by native speakers.

2 As mentioned in the beginning, there are more studies that offer several languages and where the respondent can choose the language. However, one problem of the data analysis of other studies that should not be underestimated is that sometimes the language used is not logged at all, or at least not listed in scientific use files, or the respondent's mother tongue is not surveyed. This once again illustrates the lack of attention given to the possible influence of language on the data.

In total, 74.10% of the screens analyzed were answered by the respondents in their mother tongue. However, these screens with language matches are distributed differently among the respondents. A total of 64.34% of respondents conducted the entire survey in their mother tongue and thus had a language match for each question, whereas 21.66% did not answer a single question in their mother tongue. This already indicates that only a small proportion of respondents took advantage of the opportunity to change languages during the survey. In fact, looking at the number of times respondents changed languages, 78.18% of respondents did not change language at all (regardless of whether the language used was their mother tongue or not) and less than 4 % changed the language more than 10 times.

A parent interview in this first wave contained just over 300 questions and an adolescent interview contained at least 250 questions. However, not all items were considered. Items for which translation quality problems were identified through the translation process of follow-up waves and for which the translation was therefore modified in the follow-up waves are excluded in the corresponding language. With all these limitations, the analysis sample contained 1,865 persons who together answered a total of 227,448 items. Due to the restrictions made by operationalization (see below), for 219,325 items, it was possible to clearly identify the language in which the item was answered (although this number and the number of respondents will be lower in the multivariate analyses due to missing values in the data).[3] This is also the basis for Table 1.

Operationalization

The independent variable of whether an item was viewed in a native or non-native language is no longer clearly identifiable in the case of items where the respondent changed the language several times. Therefore, only items that were either viewed in only one language or for which the language was switched only one time by the respondents were considered. In the latter cases, it is assumed that the item was answered in the language that was switched to. In Hypothesis 3, in which the theoretically assumed explanation for the respondent's behavior is language-dependent memory, only items that were viewed in one language without changing language were considered to ensure that the memories were associated with only one language.

The time for answering an item in Hypothesis 1 is measured as the respondent's cumulative time spent viewing an item. All items viewed by a respondent

3 For detailed descriptive analyses of the complete sample of the ReGES study see Will et al. (2018) and Appendix 1.

for more than ten minutes[4] were considered interview interruptions and were not included in the analyses.

To measure the accuracy of open answers in Hypothesis 2, all open numerical answers except dates are considered. Dates are not considered because, on the one hand, dates such as month and year of birth or month and year of schooling are easier to remember and therefore less cognitively demanding (Burton & Blair, 1991); on the other hand, cultural differences[5] have to be considered. For these reasons, only open numerical answers that refer to frequencies are considered.[6] In each questionnaire, there was a maximum of 7 open numerical answers. As done in comparable research (e.g., Holbrook et al., 2014; Schober et al., 2015), it is expected that all answers divisible by 5 will tend to be rounded so that the accuracy of open answers is operationalized in binary form: If the answer is divisible by 5, it is assumed to be rounded and therefore less accurate (*1 = rounded*).

To test Hypothesis 3, items on religiosity and gender role attitudes are analyzed because it can be assumed here that social norms differ between Syria, the country of origin, and Germany, the host country. It can be assumed that gender roles are more traditional in Syria than in Germany and that these gender roles are anchored in the mother tongue of the respondents. If, however, items about gender roles (using a sum score from 4 *"egalitarian"* to 16 *"traditional"*) or religiosity (using a four-point ordinal scale from *"not at all religious"* to *"very religious"*) are answered in a non-native language, it can be assumed that the question will be answered more rationally and will probably be answered in a more socially desirable way, e.g., in accordance with the norms of the host country. In the case of Germany as the host country, therefore, the answer will be more liberal or secular.

Methods of Analysis

To test the hypotheses, different regression analyses are performed for each hypothesis. It is important to remember that the objective of the ReGES study was to describe educational trajectories of refugees. Consequently, the data

4 It would also be reasonable to evaluate an interruption at 5 minutes or at 15 minutes; however, the results presented would not differ in the core of the conclusions.

5 This is less about different calendars than, for example, the phenomenon that the target group of the ReGES study is born in January more often than average. One potential explanation for this phenomenon is that in some cultures, birthdays are not a significant event and are therefore not celebrated. Therefore, some refugees may be unaware of their birthday, resulting in the mention of 01.01. as the birthday in official documents to avoid leaving the date and month empty.

6 An overview of these 13 items can be found in Appendix 2. However, since these 13 items also include many items that were asked separately for each child in the parent questionnaire, the number of items actually asked varies greatly.

were not collected through an experimental design.[7] Therefore, possible confounding variables are included in the analyses: Whether an item is answered in a mother tongue or not can be influenced by the country of origin (e.g., the national language of the country of origin is not offered as the language of the survey), by the length of stay in Germany (e.g., longer stay in Germany improves German language skills and thus the probability that items are answered in German), by education (e.g., higher education usually means better non-native language skills and thus higher chances to answer items in a non-native language) and age (e.g., cohort effect: today, non-native languages are taught more often at school, so that young people are more likely to be able to answer items in a non-native language). All these variables could also affect the dependent variable of the respective hypothesis in different ways.[8]

Additionally, a translation issues may result in respondents preferring to answer items in a language other than their mother tongue. Although all items for which translation problems were identified (when using them in later waves) were excluded from the analyses, regional discrepancies in item comprehension may persist, particularly in Arabic and Kurmanji (Gentile et al., 2019). Furthermore, it is also possible that the language was changed for one of the items identified as having been translated inaccurately and that the language was not changed back for the next items (which are included in the analyses). Additionally, unidentified translation issues can increase the cognitive load and make the understanding of an item more difficult. Therefore, the translation is additionally included as a control variable in Hypotheses 1 and 2, which are based on cognitive load and item understanding.

In Hypothesis 1, the length of an item, as measured by the number of characters, may affect respondents' preference to read longer texts in their native language. At the same time, it can be assumed that the length of a text also influences the time taken to answer an item. Therefore, this variable is also included in Hypothesis 1 as a control variable. When examining all possible screens that could be displayed in the CASI questionnaire, a quantitative analysis reveals that the average character lengths of the question in its original German version is

7 Obviously, languages were not randomly assigned. However, it can be assumed that the selection of language depends less on the characteristics of the respondents and more on the individual interview situation. An analysis of language changes within the analyzed CASI and across the follow-up interviews (that were not completely self-administered) shows, for example, that the selected language remained constant for less than 25% of the parents analyzed here. Examples of these situational factors are the availability of interviewers in the respondent's mother tongue or comprehension problems due to translation issues.

8 Furthermore, the proficiency in different languages may also influence the language in which an individual prefers to respond, as well as linguistic comprehension and the cognitive load associated with language processing. Thus, language competence may be an additional confounding variable. Unfortunately, this variable is not included in the ReGES study for all languages, which must be considered when interpreting the results.

239. The mean length of the texts of the questions in English was slightly shorter (223 characters) and considerably longer in Arabic with 312 characters.

Except for the two last-mentioned variables, all other control variables are characteristics of the respondents and not of the item, which are the actual units of analysis. Thus, the items are nested in a two-level structure by respondents, and standard errors clustered by respondents are estimated in the regression analyses. Furthermore, since the proportion of men and women in the sample is clearly biased[9], gender is also included as a control variable to avoid sample bias.

Empirical Results

As discussed in the previous section, different regression models are used in Table 2 for the multivariate analyses. In Table 2, Model 1.1, using a linear regression model (OLS model, robust standard errors clustered 1,490 respondents), the significant coefficient shows that items answered in a mother tongue took an average of 1.17 seconds longer. A possible explanation for this could be that, on the one hand, Arabic was the native language in 90.99% of the items that were answered in Arabic. Arabic items have more characters on average than German or English, which more often represent non-native languages. Keeping the control variables constant, the highly significant coefficient shows that respondents need an average of 2.44 seconds less time to answer an item in their mother tongue than an item in a non-native language. Therefore, the results in Model 1.2 support H1.

In contrast, H2 cannot be confirmed based on the analyses. The coefficient of the binary logistic regression (average marginal effects, robust standard errors clustered for 1,461 respondents) in Model 2.2 is not significant after including the control variables, even at a significance level of 10%.

The linear regression (OLS model) in Model 3.1 in Table 2 confirms the assumption of H3 and shows that items are answered more traditionally in a mother tongue than in a non-native language. The magnitude and significance of this coefficient increases when the control variables (in Model 3.2) are included so that H3 can be confirmed by analyzing gender roles. A separate analysis of the models for adolescents and parents (see Appendix 3) shows that, including the control variables, the effect is greater for parents with a coefficient of 1.72, while

9 Although this imbalance corresponds to the gender distribution of the refugees in Germany (e.g., Neske & Rich, 2016; Rich, 2016), given that families are interviewed in the ReGES study, it can be assumed here (and the feedback from the interviewers has also shown) that fathers as "classical heads of household" are more likely to answer the CASI interview than women so that the sample seems to be slightly self-selective.

Table 2 Multivariate analyses

	Hypothesis 1		Hypothesis 2		Hypothesis 3 (gender roles)	
Dependent variable	Duration in seconds		Accuracy (rounded = 1)		Gender roles, 4 (egalitarian) to 16 (traditional)	
	Model 1.1	Model 1.2	Model 2.1	Model 2.2	Model 3.1	Model 3.2
Item in mother tongue	1.17* (0.47)	-2.44*** (0.67)	-0.06* (0.02)	0.02 (0.04)	0.76** (0.25)	0.87*** (0.25)
Controls:						
Country of origin		✓		✓		✓
Age		✓		✓		✓
Length of stay in Germany		✓		✓		✓
Education		✓		✓		✓
Gender		✓		✓		✓
Translation		✓		✓		
Length of text		✓				
Pseudo/ Adjusted R^2	.000	.011	.003	.021	.007	.033
N	168,459	168,459	2,667	2,667	1,143	1,143

Source: ReGES data, own calculations, Wave 1.
Notes: Estimates with standard errors in parentheses.
* $p < .05$,
** $p < .01$,
*** $p < .001$.

the effect is not significant for adolescents. One possible explanation for this is that gender roles are not yet as pronounced in adolescents and are therefore less linked to the mother tongue.

Table 3 shows that the results of the ordinal logistic regression (average marginal effects) with religiosity as the dependent variable correspond to H3. It can be assumed that most people tend to experience religion in their mother tongue, so that the level of religiosity is reported to be higher in the mother tongue than in a non-native language. Even when controlling for variables such as age, education or length of stay in Germany, the significant average marginal effects show that, for example, the probability that a person states that he or she is very religious is almost 3 percentage points higher in the mother tongue than if the

Table 3 Ordinal logistic regression for religiosity (average marginal effects)

	Model 1	Model 2
Item only in mother tongue		
Not at all religious	-0.03**	-0.03**
	(0.01)	(0.01)
Not very religious	-0.07**	-0.06**
	(0.02)	(0.02)
Quite religious	0.08**	0.07**
	(0.03)	(0.03)
Very religious	0.03***	0.03**
	(0.01)	(0.01)
Controls:		
Age, length of stay in Germany, gender, education		✓
Adjusted R^2	.004	.018
N	1,377	1,377

Source: ReGES data, own calculations, Wave 1.
Notes: Estimates with standard errors in parentheses.
* $p < .05$,
** $p < .01$,
*** $p < .001$.

item was answered in a non-native language. In line with the analysis about gender roles, the effect in Model 2 is also stronger for the parents and no longer significant for the adolescents in separate analyses (see Appendix 4).

Discussion

Due to the increasing number of migrants, there are more and more people who have a mother tongue other than the national language. At the same time, technological innovation has made it possible for respondents in many surveys to choose whether to answer questions in their native language or in a non-native language. The present analyses based on data from the German refugee study ReGES have shown that there is a difference in response behavior when a question is answered in a native language instead of a non-native language, even if not all hypotheses could be confirmed. The data have shown that when considering the time for answering an item, the cognitive load seems to be higher and the understanding of a question is more difficult when a question is presented and answered in a non-native language (H1). Depending on how long a survey is,

this effect can be clearly noticeable in the overall time and thus possibly cause exhaustion or reduce the respondents' willingness to cooperate in subsequent follow-up surveys of a panel study.

A relationship between more precise information and language nativeness (H2) could not be demonstrated. However, since the first model for Hypothesis 2 without control variables shows a significant effect, it can be assumed that both the willingness for higher cognitive effort and to give accurate unrounded information and the willingness to answer an item in a non-native language that is difficult to understand depend strongly on the control variables included (such as age, gender or on the cognitive willingness or ability to perform in general, for which the educational level can be regarded as an indicator).

The significant coefficients in Model 2.2 and 3.2 (Table 2) show a clear correlation between the language used and responses to sensitive questions. However, regarding causality, a few limitations of the data and the study design of the ReGES study must be considered here. It is possible that an individual's religiosity or attitudes toward gender roles may also influence the choice of language. For example, more liberal respondents might be more willing to answer a survey in a non-native language. However, field experience indicates that the selection of language is more dependent on the interview situation. For instance, an analysis of the language used by parents within the CASI of the first survey wave combined with the starting language used in the follow-up interviews that were not fully self-administered reveals—without considering language changes within the interviews of the follow-up waves—that the language used remained consistent for less than 25% of the parents.[10] Furthermore, the fact that respondents have no rational reasons for voluntarily conducting the interview in a non-native language instead of their native language (according to the assumptions of RCT and cognitive load theory) strengthens the hypothesis that external, situational factors and peculiarities of the ReGES study are responsible for the fact that respondents did not complete the survey in their mother tongue, although this was offered in almost all cases. The data showed—similar to other studies (e.g., Kinnunen et al., 2015)—that even if the mother tongue is offered, the mother tongue is not automatically chosen. An investigation of these factors would provide more evidence (and would also help to evaluate whether the costs and efforts for a multilingual survey are truly worth it). In order to conduct a

10 It is clear that the refugees' German language abilities will continue to develop over time, allowing for an increasing number of respondents to be interviewed in German in subsequent survey waves. However, if the language selected is dependent on the characteristics of the respondents (e.g., religiosity), it can be reasonably assumed that the language used by the respondents will remain stable throughout the interview and across survey waves. This indicates that situational factors (e.g., availability of native speaker interviewers, comprehension issues, etc.) may be more influential than respondents' self-selection in determining language choice.

more detailed analysis of the effects of the language used, future survey experiments can be used to confirm and understand the aforementioned effects.

Furthermore, the individual relational explanatory power of the language match between the respondent's mother tongue and the language used with a goodness of fit R^2 of less than .05 in each analysis is quite low in all the models presented, so it can be assumed that there are many more aspects that contribute to explaining different response behavior. For example, cultural considerations as well as linguistic characteristics of a language (such as the possible influence of grammatical gender (e.g., Boroditsky et al., 2003; Garnham et al., 2016)) could also help to explain language-dependent response behaviors. Regarding linguistic aspects, another point that would be relevant for further research would be whether and to what extent dialects also cause a difference in the response to items. According to language-dependent memory theory, this should have an influence, which would be relevant for the language focused on in this paper, Arabic. Indeed, it can be assumed that not all respondents who stated that Arabic is their mother tongue actually learned the Modern Standard Arabic used in the questionnaire as their mother tongue but rather an Arabic dialect.

Conclusion

The results of the influence of language show that language is relevant to response behavior and needs to be taken into account from a survey methodological point of view. This not only means that more research is to be done on this topic but also that the complexity of non-native-language surveys requires more attention in the practical implementation of surveys.

Even though one hypothesis could not be confirmed, and thus only a few practice-relevant statements on the effect of the language used in surveys can be made, the present paper shows that it can make a difference whether an item was answered in a mother tongue or in a non-native language. Therefore, the language used as a possible factor in response behavior should not be neglected, especially since an increasing number of surveys are offered in multiple languages.

Both the empirical results and the theoretical considerations suggest that it makes sense to offer surveys for respondents in their mother tongue. It enables people to participate who otherwise would not have been able to due to language barriers (e.g., Feskens et al., 2006; Jacobsen, 2018). Especially in surveys where, according to rational choice theory, there are hardly any incentives to take part, offering their mother tongue enables the respondents to participate with less cognitive load, which might positively influence the motivation and thus the data quality. Furthermore, offering surveys in more mother tongues might—which should be investigated additionally (Watson & Wooden, 2009, p. 165)—influence the general willingness to cooperate by showing respect for the respondent.

Data

Refugees in the German Educational System (2021). Raw data. Leibniz Institute for Educational Trajectories. Scientific-Use-File available via https://doi:10.5157/ReGES:RC1:SUF:1.0.0 and https://doi.org/10.5157/ReGES:RC2:SUF:1.0.0

(Note: the data on which language was used for which item is currently not published for reasons of data protection. Therefore, the raw data was used. In the raw data there is one variable of which the recoding is also explained in the do-files.)

References

Battistin, E., Miniaci, R., & Weber, G. (2003). What do we learn from recall consumption data? *The Journal of Human Resources, 38*(2), 354–385. https://doi.org/10.2307/1558748

Blohm, M., & Diehl, C. (2001). Wenn Migranten Migranten befragen: Zum Teilnahmeverhalten von Einwanderern bei Bevölkerungsbefragungen. *Zeitschrift für Soziologie, 30*(3), 223–242. https://doi.org/10.1515/zfsoz-2001-0304

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 61–78). Bradford Books. https://doi.org/10.7551/mitpress/4117.003.0010

Burton, S., & Blair, E. (1991). Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly, 55*(1), 50–79. https://doi.org/10.1086/269241

Calderón, J. L., Morales, L. S., Liu, H., & Hays, R. D. (2006). Variation in the readability of items within surveys. *American Journal of Medical Quality, 21*(1), 49–56. https://doi.org/10.1177/1062860605283572

Caldwell-Harris, C. L., & Ayçiçeği-Dinn, A. (2009). Emotion and lying in a non-native language. *International Journal of Psychophysiology, 71*(3), 193–204. https://doi.org/10.1016/j.ijpsycho.2008.09.006

Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The moral foreign-language effect. *Philosophical Psychology, 29*(1), 23–40. https://doi.org/10.1080/09515089.2014.993063

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS ONE, 9*(4), e94842. https://doi.org/10.1371/journal.pone.0094842

Danziger, S., & Ward, R. (2010). Language changes implicit associations between ethnic groups and evaluation in bilinguals. *Psychological Science, 21*(6), 799–800. https://doi.org/10.1177/0956797610371344

Dewaele, J.-M., & Nakano, S. (2012). Multilinguals' perceptions of feeling different when switching languages. *Journal of Multilingual and Multicultural Development, 34*(2), 107–120. https://doi.org/10.1080/01434632.2012.712133

Díaz-Morales, J. F., Ferrari, J. R., Díaz, K., & Argumedo, D. (2006). Factorial structure of three procrastination scales with a Spanish adult population. *European Journal of Psychological Assessment, 22*(2), 132–137. https://doi.org/10.1027/1015-5759.22.2.132

Dotinga, A., van den Eijnden, R. J. J. M., Bosveld, W., & Garretsen, H. F. L. (2005). The effect of data collection mode and ethnicity of interviewer on response rates and

self-reported alcohol use among Turks and Moroccans in the Netherlands: An experimental study. *Alcohol and Alcoholism, 40*(3), 242–248. https://doi.org/10.1093/alcalc/agh144

Esser, H. (1990). „Habits", „Frames" und „Rational Choice". Die Reichweite von Theorien der rationalen Wahl (am Beispiel der Erklärung des Befragtenverhaltens). *Zeitschrift für Soziologie, 19*(4), 231–247. https://doi.org/10.1515/zfsoz-1990-0401

Feskens, R., Hox, J., Lensvelt-Mulders, G., & Schmeets, H. (2006). Collecting data among ethnic minorities in an international perspective. *Field Methods, 18*(3), 284–304. https://doi.org/10.1177/1525822X06288756

Garnham, A., Oakhill, J., von Stockhausen, L., & Sczesny, S. (2016). Editorial: Language, cognition, and gender. *Frontiers in Psychology, 7*, 772. https://doi.org/10.3389/fpsyg.2016.00772

Geipel, J., Hadjichristidis, C., & Surian, L. (2015). The foreign language effect on moral judgment: The role of emotions and norms. *PLoS ONE, 10*(7), e0131529. https://doi.org/10.1371/journal.pone.0131529

Gentile, R., Heinritz, F., & Will, G. (2019). *Übersetzung von Instrumenten für die Befragung von Neuzugewanderten und Implementation einer audiobasierten Interviewdurchführung* (LIfBi Working Paper No. 86). Leibniz Institute for Educational Trajectories. https://www.lifbi.de/Portals/2/Working%20Papers/WP_LXXXVI.pdf

Gideon, M., Helppie-McFall, B., & Hsu, J. W. (2017). Heaping at round numbers on financial questions: The role of satisficing. *Survey Research Methods, 11*(2), 189–214. https://doi.org/10.18148/srm/2017.v11i2.6782

Glantz, A., & Michael, T. (2014). Interviewereffekte. In N. Baur & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 313–322). Springer Fachmedien. https://doi.org/10.1007/978-3-531-18939-0_21

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). John Wiley & Sons.

Guarnaccia, P. J., Angel, R., & Worobey, J. L. (1989). The factor structure of the CES-D in the Hispanic health and nutrition examination survey: The influences of ethnicity, gender and language. *Social Science & Medicine, 29*(1), 85–94. https://doi.org/10.1016/0277-9536(89)90131-7

Hadjichristidis, C., Geipel, J., & Boaz, K. (2019). The influence of native language in shaping judgement and choice. *Progress in Brain Research, 247,* 253–272. https://doi.org/10.1016/bs.pbr.2019.02.003

Hadjichristidis, C., Geipel, J., & Surian, L. (2019). Breaking magic: Foreign language suppresses superstition. *Quarterly Journal of Experimental Psychology, 72*(1), 18–28. https://doi.org/10.1080/17470218.2017.137178

Harris, C. L., Ayçiçeği, A., & Gleason, J. B. (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied Psycholinguistics, 24*(4), 561–579. https://doi.org/10.1017/S0142716403000286

Harris, C. L., Gleason, J. B., & Ayçiçeği, A. (2006). When is a first language more emotional? Psychophysiological evidence from bilingual speakers. In A. Pavlenko (Ed.), *Bilingual minds: Emotional experience, expression, and representation* (pp. 257–283). Multilingual Matters. https://doi.org/10.21832/9781853598746-012

Harzing, A.-W., & Maznevski, M. (2002). The interaction between language and culture: A test of the cultural accommodation hypothesis in seven countries. *Language and Intercultural Communication, 2*(2), 120–139. https://doi.org/10.1080/14708470208668081

Hasegawa, M., Carpenter, P. A., & Just, M. A. (2002). An fMRI study of bilingual sentence comprehension and workload. *NeuroImage, 15*(3), 647–660. https://doi.org/10.1006/nimg.2001.1001

Hayakawa, S., & Keysar, B. (2018). Using a foreign language reduces mental imagery. *Cognition, 173*, 8–15. https://doi.org/10.1016/j.cognition.2017.12.010

Heinritz, F., & Will, G. (2021). *Selektive Teilnahme von Geflüchteten an der Panelstudie ReGES* (LIfBi Working Paper No. 96). Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/LIfBi:WP96:1.0

Holbrook, A. L., Anand, S., Johnson, T. P., Cho, Y. I., Shavit, S., Chávez, N., & Weiner, S. (2014). Response heaping in interviewer-administered surveys: Is it really a form of satisficing? *Public Opinion Quarterly, 78*(3), 591–633. https://doi.org/10.1093/poq/nfu017

Hunt, S. M., & Bhopal, R. (2004). Self report in clinical and epidemiological studies with non-English speakers: the challenge of language and culture. *Journal of Epidemiology & Community Health, 58*(7), 618–622. https://doi.org/10.1136/jech.2003.010074

International Labour Office. (2010). *International labor migration: A rights-based approach.* International Labour Organization. https://webapps.ilo.org/public/libdoc/ilo/2010/110B09_59_engl.pdf

International Labour Office. (2018). *Ilo global estimates on international migrant workers: Results and methodology.* International Labour Organization. https://www.onlinelibrary.iihl.org/wp-content/uploads/2020/05/2018-I1-2.pdf

International Organization for Migration. (2019). *World migration report 2020*. International Organization for Migration. https://publications.iom.int/system/files/pdf/wmr_2020.pdf

Jacobsen, J. (2018). Language barriers during the fieldwork of the IAB-BAMF-SOEP survey of refugees in Germany. In D. Behr (Ed.), *Surveying the migrant population: Consideration of linguistic and cultural issues* (pp. 75–84). GESIS – Leibniz-Institut für Sozialwissenschaften. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-58536-7

Kappelhof, J. (2017). Survey research and the quality of survey data among ethnic minorities. In P. P. Biemer, E. D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 235–252). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119041702.ch11

Kinnunen, J. M., Malin, M., Raisamo, S. U., Lindfors, P. L., Pere, L. A., & Rimpelä, A. H. (2015). Feasibility of using a multilingual web survey in studying the health of ethnic minority youth. *JMIR Research Protocols, 4*(2), e53. https://doi.org/10.2196/resprot.3655

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lee, T. (2002). *Language-of-interview effects and Latino mass opinion*. SSRN. https://doi.org/10.2139/ssrn.303165

Loosveldt, G. (2008). Face-to-face interviews. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 201–220). Taylor & Francis Group/Lawrence Erlbaum Associates.

Marian, V., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology, 20*(8), 1025–1047. https://doi.org/10.1002/acp.1242

Marian, V., & Kaushanskaya, M. (2004). Self-construal and emotion in bicultural bilinguals. *Journal of Memory and Language, 51*(2), 190–201. https://doi.org/10.1016/j.jml.2004.04.003

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General, 129*(3), 361–368. https://doi.org/10.1037/0096-3445.129.3.361

Marin, G., Triandis, H. C., Betancourt, H., & Kashima, Y. (1983). Ethnic affirmation versus social desirability. *Journal of Cross-Cultural Psychology, 14*(2), 173–186. https://doi.org/10.1177/0022002183014002003

McKay, R. B., Breslow, M. J., Sangster, R. L., Gabbard, S. M., Reynolds, R. W., Nakamoto, J. M., & Tarnai, J. (1996). Translating survey questionnaires: Lessons learned. *New Directions for Evaluation, 70*, 93–104. https://doi.org/10.1002/ev.1037

Neske, M., & Rich, A.-K. (2016). *Asylerstantragsteller in Deutschland im ersten Halbjahr 2016: Sozialstruktur, Qualifikationsniveau und Berufstätigkeit* (BAMF-Kurzanalyse 4-2016). Bundesamt für Migration und Flüchtlinge. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-67527-7

Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H.-Y. (2016). Rational learners and moral rules. *Mind & Language, 31*(5), 530–554. https://doi.org/10.1111/mila.12119

Paas, F., & Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*(4), 351–371. https://doi.org/10.1007/BF02213420

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8

Pan, Y., Leeman, J., Fond, M., & Goerman, P. (2014). *Multilingual survey design and fielding: Research perspectives from the U.S. census bureau* (Research Report Series 2014-01). U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/rsm2014-01.pdf

Pérez, E. O. (2009). Lost in translation? Item validity in bilingual political surveys. *The Journal of Politics, 71*(4), 1530–1548. https://doi.org/10.1017/S0022381609990156

Peytcheva, E. (2018). Can the language of survey administration influence respondents' answers? In T. P. Johnson, B.-E. Pennell, I. A. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts* (pp. 325–340). John Wiley & Sons, Inc.

Rich, A.-K. (2016). *Asylerstantragsteller in Deutschland im Jahr 2015: Sozialstruktur, Qualifikationsniveau und Berufstätigkeit* (BAMF-Kurzanalyse 3-2016). Bundesamt für Migration und Flüchtlinge. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-67504-2

Schober, M. F., Conrad, F. G., Antoun, C., Ehlen, P., Fail, S., Hupp, A. L., Johnston, M., Vickers, L., Yan, H. Y., & Zhang, C. (2015). Precision and disclosure in text and voice interviews on smartphones. *PLoS ONE, 10*(6), e0128337. https://doi.org/10.1371/journal.pone.0128337

Schoua-Glusberg, A. (2004). *Assessing comprehension of translated questionnaires with qualitative methods* (Statistics Canada International Symposium Series – Proceedings No. 11-522-X20040018746). Statistics Canada. https://www150.statcan.gc.ca/n1/pub/11-522-x/2004001/8746-eng.pdf

Shin, H. I., & Kim, J. (2017). Foreign language effect and psychological distance. *Journal of Psycholinguistic Research, 46*(6), 1339–1352. https://doi.org/10.1007/s10936-017-9498-7

Steinhauer, H.-W., Zinn, S., & Will, G. (2019). Sampling refugees for an educational longitudinal survey. *Survey Methods: Insights from the Field*. Advance online publication. https://doi.org/10.13094/SMIF-2019-00007

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73–100). The National Academic Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–83. https://doi.org/10.1037/0033-2909.133.5.859

von Maurice, J., & Will, G. (2023). Data from the panel study 'Refugees in the German Educational System (ReGES)'. *Journal of Open Psychology Data, 11*(1), 1–15 https://doi.org/10.5334/jopd.77

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 157–81). Wiley. https://doi.org/10.1002/9780470743874.ch10

Welch, S., Comer, J., & Steinman, M. (1973). Interviewing in a Mexican-American community: An investigation of some potential sources of response bias. *Public Opinion Quarterly, 37*(1), 115–126. https://doi.org/10.1086/268065

Will, G., Balaban, E., Dröscher, E., Homuth, C., & Welker, J. (2018). *Integration von Flüchtlingen in Deutschland: Erste Ergebnisse aus der ReGES-Studie* (LIfBi Working Paper No 76). Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/LIfBi:WP76:2.0

Will, G., Homuth, C., von Maurice, J., & Roßbach, H.-G. (2021). Integration of recently arrived underage refugees: Research potential of the study ReGES—Refugees in the German Educational System. *European Sociological Review, 37*(6), 1027–1043. https://doi.org/10.1093/esr/jcab033

Appendix

Appendix 1 Characteristics of control variables on the level of respondents

	M/Freq.	*SD*	Min.	Max.
Country of origin				
Afghanistan	6.38 %			
Iraq	8.69 %			
Iran	1.82 %			
Syria	77.05 %			
Other	6.06 %			
Age	27.84	13.90	14	75
Length of stay in Germany	28.81	9.20	3	53
Gender				
Male	62.47 %			
Female	37.53 %			
Education				
ISCED 0: Preprimary education	4.74 %			
ISCED 1: Primary Education	20.48 %			
ISCED 2: Lower secondary education	11.26 %			
ISCED 3: Upper secondary education	17.05 %			
ISCED 4: Postsecondary nontertiary education	4.24 %			
ISCED 5: Short-cycle tertiary education	6.65 %			
ISCED 6: Bachelor or equivalent	7.77 %			
ISCED 7: Master of equivalent	8.10 %			
ISCED 8: Doctoral or equivalent	0.16 %			
Missing values	19.57 %			

Source: ReGES data, own calculations, Wave 1, n = 1,865 respondents.

Appendix 2 Items used for measuring accuracy

Name	Question
p3100000	The following questions are about your living situation in Germany now. In how many different accommodation facilities have you lived since your arrival in Germany? Please list all stations from the preliminary reception center to your current accommodation.
p3241140	On average, how many hours does your child spend at the childcare facility per week?
p6242120	How many hours of German language classes does your child attend at preschool per week?
p6242140	How many hours of German language classes does your child attend outside of his or her preschool per week?
p3140000	The following questions are about your child's living situation in Germany. In how many different accommodation facilities has your child lived since his or her arrival in Germany? Please list all stations from the preliminary reception center to your current accommodation.
p6242220	How many hours of German language classes does your child attend per week?
p3241180	On average, how many hours per week does your child spend with a childminder or nanny?
p6242410	On a normal weekday, how many hours does your child spend in situations where he or she hears or speaks German?
p3241250	On average, how many hours did your child spend at the childcare facility per week?
p3241300	On average, how many hours per week did your child spend with a childminder or nanny?
t6242220	For how many hours a week do you take German classes for refugees and migrants at school?
t6242240	For how many hours a week do you take German classes for refugees and migrants outside of school?
t6242420	On a normal weekday, how many hours do you spend in situations where you hear, speak, read or write German?

Source: ReGES, parents- and adolescent questionnaires, Wave 1.

Appendix 3 Multivariate analyses, separately for adolescents and parents

	Hypothesis 1		Hypothesis 2		Hypothesis 3 (gender roles)	
Dependent variable	Duration in seconds		Accuracy (rounded = 1)		Gender roles, 4 (egalitarian) to 16 (traditional)	
	Model 1.1	Model 1.2	Model 2.1	Model 2.2	Model 3.1	Model 3.2
	Parents	Adolescents	Parents	Adolescents	Parents	Adolescents
Item in mother tongue	-2.28** (0.80)	-2.87* (1.20)	0.90 (0.24)	2.12 (1.06)	1.72*** (0.32)	-0,56*** (0.41)
Controls:						
Country of origin	✓	✓	✓	✓	✓	✓
Age	✓	✓	✓	✓	✓	✓
Length of stay in Germany	✓	✓	✓	✓	✓	✓
Education	✓	✓	✓	✓	✓	✓
Gender	✓	✓	✓	✓	✓	✓
Translation	✓	✓	✓	✓		
Length of text	✓	✓				
Pseudo/ Adjusted R^2	.015	.008	.034	.036	.052	.037
N	99,069	69,390	1,903	764	729	414

Source: ReGES data, own calculations, Wave 1.
Notes: Estimates with standard errors in parentheses.
* $p < .05$,
** $p < .01$,
*** $p < .001$.

Appendix 4 Ordinal logistic regression for religiosity (average marginal effects, Model 2), separately for adolescents and parents

	Parents	Adolescents
Item only in mother tongue		
Not at all religious	-0.05**	0.00
	(0.02)	(0.02)
Not very religious	-0.11***	0.00
	(0.03)	(0.03)
Quite religious	0.12**	0.00
	(0.04)	(0.03)
Very religious	0.04***	0.00
	(0.01)	(0.02)
Controls:		
Age, length of stay in Germany, gender, education	✓	✓
Adjusted R^2	.025	.023
N	848	529

Source: ReGES data, own calculations, Wave 1.
Notes: Estimates with standard errors in parentheses.
* $p < .05$,
** $p < .01$,
*** $p < .001$.

From Clicks to Quality: Assessing Advertisement Design's Impact on Social Media Survey Response Quality

Jessica Donzowa[1, 2], Simon Kühne[2] & Zaza Zindel[2, 3]

[1] *Max Planck Institute for Demographic Research*
[2] *Bielefeld University*
[3] *German Center for Integration and Migration Research (DeZIM)*

Abstract

Researchers are increasingly using social media platforms for survey recruitment. However, empirical evidence remains sparse on how the content and design characteristics of advertisements used for recruitment affect response quality in surveys. Building on leverage-salience and self-determination theory, we assess the effects of advertisement design on response quality. We argue that different advertisement designs may resonate with specific social groups who vary in their commitment to the survey, resulting in differences in the observed response quality. We use data from a study conducted via ads placed on Facebook in Germany and the United States in June 2023. The survey, focusing on attitudes toward climate change and immigration, featured images with varying thematic associations with the topics (strong, loose, neutral). The sample consisted of 4,170 respondents in Germany and 5,469 respondents in the United States. We compare several data quality indicators, including break-off rate, completion time, non-differentiation, item non-response, passing an attention check question, and follow-up availability, across different advertisement features. Regression analyses indicate differences in response quality across advertisement designs, with a strong thematic design generally being associated with poorer response quality. Strongly themed ad designs are generally associated with higher attrition, non-differentiation, and item non-response, and with a lower probability of passing an attention check and providing an e-mail address for future survey inquiries. Our study advances the literature by highlighting the substantial impact of advertisement design on survey data quality, and emphasizing the importance of tailored decision-making in recruitment design for social media-based survey research.

Keywords: social media recruitment, advertisement design, online survey, survey topic interest, response quality, survey invitation design

The use of social media for (online) survey recruitment has grown over the past decade, with the majority of researchers using Facebook by Meta Inc. as a recruitment tool (Zindel, 2023). Although initially designed for business purposes, research has demonstrated that Meta's advertisement manager is effective for recruiting online survey participants (Grow et al., 2022; Iannelli et al., 2020; Kühne & Zindel, 2020; Pötzschke & Braun, 2017). Similar to companies that use advertisements to promote their services and products, researchers can use advertisements—an image or a video with some text and a link—to recruit respondents. Thus, ads on social media represent a form of digital survey invitation that, unlike email or postal invitations, centers around images or videos to draw users' attention on multi-media platforms. At the same time, concerns about data quality remain, particularly regarding representation (e.g., self-selection biases and non-representativeness) and measurement error (e.g., satisficing behavior due to the lack of interviewer presence; De Man et al., 2021; Heerwegh & Loosveldt, 2008).

Previous studies examining the design effects of ad-based survey invitations in social media have largely focused on response and break-off rates (e.g., Choi et al., 2017, Stern et al., 2022). However, there is almost no empirical evidence on how advertisement properties affect the response quality in Facebook-recruited surveys beyond participation. This is somewhat surprising given the vast literature on the design effects (of invitations) on response quality for mail or web surveys (e.g., Haer & Meidert, 2013; Kaplowitz et al., 2012; Keusch, 2013; Mavletova et al., 2014).

Data quality assessments in the social sciences remain fragmented, and there is a need for systematic frameworks to assess different dimensions of data quality (Birkenmaier et al. 2024). This paper aims to contribute to the broader discussion of data quality by focusing specifically on the intrinsic requirements of social media-recruited survey data, particularly the risk of measurement error. We designed a study that varied images in advertisements used for survey recruitment on Facebook in Germany and the United States in 2023. In addition to a "neutral" set of images, we tested images with varying degrees of association with two survey topics: immigration and climate change. In the analyses, we estimated the effects of these topics and ad image properties on several data

Acknowledgements

Jessica Donzowa gratefully acknowledges the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS). This study was funded with support from the Max Planck Institute for Demographic Research, which is part of the Max Planck Society.

Direct correspondence to

Jessica Donzowa, Max Planck Institute for Demographic Research (MPIDR), Rostock, Germany & Bielefeld University, Bielefeld, Germany
E-mail: donzowa@demogr.mpg.de

quality indicators, including survey break-off rate, speeding behavior, non-differentiation, item non-response, attentiveness, and willingness to participate in future surveys. Our approach contributes to the current state of research by a) implementing a study design that deliberately varies ad image properties across survey topics, b) focusing on the general online population in two countries (Germany and the United States), rather than on specific sub-populations, and c) testing a large set of data quality indicators. Thus, this paper is the first comprehensive study of the effects of advertisement design on response quality in surveys recruited via social media.

Background and State of Research

The use of social media—mostly Facebook—for (online) survey recruitment has steadily increased over the past decade (Zindel, 2023). While this method extends the reach of surveys, the methodological implications and the potential biases of social media recruitment are not fully understood (Lehdonvirta et al., 2021). Known biases include skewed sample compositions that favor certain populations, which can affect the reliability and validity of the resulting data (Neundorf & Öztürk, 2023). However, beyond the sample composition, the quality of the responses provided also has an impact on study outcomes.

Figure 1 Advertisement used to recruit respondents via Facebook in the United States. Desktop view.

Visual advertisement design is crucial for social media recruitment (Kühne & Zindel, 2020; Neundorf & Öztürk, 2022). Advertisements on these platforms often rely on visual elements, such as images (rarely videos), which typically make up the majority of an ad's display (see Figure 1). Because the number of texts is limited to just a few lines, the visual components often capture the initial attention of potential survey respondents and establish an initial point of engagement. In the case of image-centric ads—the most common approach in social media recruitment—a key decision regarding about what to display in an image (or multiple images) needs to be made by the researcher. Naturally, the question arises of whether the survey topic is supposed to be reflected in the images, and if so, to what extent. Alternatively, researchers have used neutral images that reflect surveys or public opinion more generally (e.g., by displaying a question mark or speech bubbles).

Existing survey methodological theories and frameworks point to several potential mechanisms through which advertisement design—here: the extent to which a survey topic is displayed in an image—can affect response quality.

The impact of survey recruitment materials—such as the design of a social media ad—on respondents' participation decision and commitment levels can be conceptualized based on the leverage-salience theory (Groves, 1992). Leverage refers to the importance of a feature of an advertisement, such as the image or the topic presented to a potential respondent. Salience, on the other hand, refers to how noticeable or prominent that feature is during the survey invitation process. Images, being the most prominent part of an ad, generally have a high degree of salience. In this context, a user's likelihood of participating in the survey is influenced by the perceived benefits of participation, and by how salient those benefits are made in the ad. Whether a feature of an ad is perceived as beneficial varies by individual characteristics, since people may perceive the same factors as more or less beneficial. Some images may resonate more with specific social groups, such as men, people with lower levels of education or specific political interests (Neundorf & Öztürk, 2022). Since different social groups are expected to show varying levels of interest in the ad, a given ad may appeal to potential respondents differently depending on their cognitive abilities, and their willingness to conscientiously participate in an online survey (Zillmann et al., 2014). Therefore, we argue that different ad designs may not only result in different sample compositions (e.g., in terms of socio-demographics), but also in different response behaviors, and, consequently, in varying levels of data quality.

Beyond socio-demographics and cognitive skills, different ad images can influence the sample composition and data quality with respect to individual motivation. Self-determination theory (Wenemark et al., 2011) not only conceptualizes the decision to participate in a survey, but also distinguishes different levels of commitment to the survey. Self-determination theory distinguishes

between autonomous extrinsic motivation, which suggests that participants are motivated by contributing to societal knowledge, and intrinsic motivation, which implies that participants find the task itself such as the survey activity enjoyable. Higher commitment, which is associated with the desire to perform the task well, is generally associated with higher response quality (Wenemark et al., 2011). Because levels and types of commitment have been associated with response quality, ad images are expected to affect the responses' quality by appealing to different motivational types through varying ad content. For example, ad images that clearly reveal that the survey topic is on a highly discussed issue, such as climate change, may systematically attract individuals with autonomous extrinsic motivation, who are motivated to contribute to climate change research. In contrast, neutral images may be more likely to attract intrinsically motivated individuals who enjoy the task of responding to surveys in general, regardless of the specific topic. In summary, leverage-salience theory and self-determination theory suggest that the characteristics and the content of ad images may not only affect the sample characteristics in terms of socio-demographics, but also systematically affect respondents' commitment, motivation, and conscientiousness in completing an online survey. Therefore, the design of ad images is expected to impact response quality.

Previous research on online surveys has shown that interest in the survey topic and the perceived burden of participation influence response quality, with greater interest in the topic being associated with higher response quality (Galesic, 2006). Conversely, an increasing desire to abandon the survey due to the response burden is reflected in decreased quality, which is particularly evident just before respondents drop out through increased item non-response (Galesic, 2006). This behavior, which is also referred to as "satisficing" or "short-cutting," refers to the tendency of respondents to choose easier response strategies to minimize their effort and individual survey burden, which may compromise data quality (Krosnick, 1999). Satisficing can include behaviors such as choosing no response (i.e., item non-response), repeating the same answers across various questions (i.e., non-differentiation or "straightlining"), and consistently agreeing with survey items (i.e., acquiescence or "yes-saying"). Research suggests that satisficing behavior is particularly likely to occur when the task difficulty is high and the respondent motivation is low (Holbrook et al., 2003; Kaminska et al., 2010; Roberts et al., 2019).

Despite the prevalence of social media recruitment, our understanding of how advertisement design influences response quality remains limited. Choi et al. (2017) found that the choice of image and wording had an impact on men's engagement levels and time spent on a mental health survey. Similarly, Stern et al. (2022) found that among young men from sexual minorities in the United States, advertisements with images resulted in fewer non-substantive survey responses than ads that used video as visual element, although the response

rates were comparable. Neundorf and Öztürk (2022) examined the effects of incentive-based versus thematic advertisements in Turkey. They found that the differences in attentiveness (passing an attention check question by checking the option "Do not know") disappeared when controlling for demographic characteristics. However, while the respondents recruited through incentive-based advertisements were more likely to answer to open-ended questions, they left shorter responses than the participants recruited through the thematic advertisements. Finally, the participants recruited through incentive-based advertisements were more likely to participate in a follow-up survey when contacted again (Neundorf & Öztürk, 2022). Donzowa et al., (2023) showed that during the early stages of the COVID-19 pandemic, more explicit survey topic display was associated with higher numbers of link clicks and higher completion rates. These results suggest that while there are economic advantages to making survey topics more explicit, the impact on response quality remains uncertain.

The theories outlined above describe the pathway through which ad design affects response quality. However, the direction of this effect is still unknown. Following the leverage-salience theory, one could argue that "thematic" advertisements with a more explicit topic presentation tend to recruit more thematically motivated respondents with a strong commitment to the survey topic, resulting in higher response quality. On the other hand, according to self-determination theory, "neutral" advertisements that make no reference to a specific survey topic may recruit respondents who are more intrinsically motivated to respond to surveys in general, resulting in consistently higher response quality. Existing research has reported inconsistent findings regarding the direction and the magnitude of this effect. Our study aims to contribute to the current state of research by implementing a design that allows us to examine the effects of ad design properties on response quality in a general online population survey setting in two countries.

Data & Methods

The following chapter presents the study design used in this project. We also present the indicators used to measure response quality and explain how they are defined.

Study Design

We use data from a survey conducted via Facebook in both Germany (June 25, 2023, to July 2, 2023) and the United States (June 25, 2023, to July 3, 2023). The survey focused on the subjects of climate change and immigration—two topics that are the subject of intense media and public debate in both countries. The

advertisement campaign used a variety of images with different thematic associations, as shown in Appendix Figure A1. Fifteen images were used in each country. Thirteen images were the same for both countries, and two images were adapted for each country. For an image showing a flag, an image of the European flag was used in Germany, and an image of the US flag was used in the United States. In the second case, one of the neutral images, images with text in the local language of each country were used ("Your opinion matters" in the United States and "Ihre Meinung ist uns wichtig" in Germany). Images were selected through a multi-step selection process. First, a broad set of images was retrieved from stock image websites (AdobeStock (https://stock.adobe.com) and iStock (https://www.istockphoto.com)), which the research team narrowed down to 73 images (43 for climate change and 30 for immigration). In a second step, these images were evaluated for their association with the survey themes through an image selection survey among the researchers' scientific network. This resulted in a set of 34 images with accompanying information about the association to the survey topic, that is, strong, loose, or neutral. In a third step, these images were used in a survey pretest based on social media recruitment via Facebook and then evaluated for their recruiting performance. The top three images for each association were selected for the final data collection.

Technically, we implemented each image in an individual campaign on Meta's advertisement management system, resulting in a total of 30 campaigns for both countries combined. The campaign's objective was to drive traffic, with the optimization goal of "link-clicks." Meta Pixel was not used in this study. From a research ethics perspective, the use of Meta Pixel may be viewed critically as it provides Meta with information about respondent behavior outside of the Facebook platform to optimize survey completion. The remaining advertising options, such as platform placement (Facebook Newsfeed) or sender of the study, were kept consistent across all ads to ensure that all campaigns had an equal chance to capture the attention of social media users and could be effectively compared. Advertisements ran exclusively on Facebook and did not include Instagram. Only demographic targeting tools were used. Geographic targeting included the respective countries, that is, regions of the U.S. and regions of Germany. The age range for all ads was set to include users of age 18 and older. We tracked through which ad a user entered our survey through Meta Ad Manager's built-in ability to define URL parameters.

When Facebook users were exposed to the ads, they had the option to self-select into the survey by clicking on the ad. Upon clicking, they were redirected to the web survey, which was hosted on Bielefeld University server and implemented using the LimeSurvey software. The web survey was optimized for mobile devices to ensure functionality and proper design across a wide range of computer and mobile device hardware and software. Prior to beginning the survey, participants were informed of the estimated length of the survey and

asked for their consent. The survey included questions about the participants' general political interests, as well as their views on immigration and climate change. Respondents were randomly assigned to begin with questions related to either immigration or climate change, irrespective of the ad image they initially encountered.

In addition to the social media recruitment efforts, the survey was replicated using a commercial panel company to allow for comparisons with other online survey populations. However, this comparison should be interpreted with caution, as the terms of participation for this survey differed from those for our social media sample, particularly in terms of compensation, as online panel respondents may be incentivized to participate. For more details on the online access panel data, see Section A2 in the appendix.

Response Quality Measures

We rely on several data quality indicators that are regularly used in the literature (i.e., survey break-off rate, speeding, non-differentiation, item non-response, passing an attention check question, and willingness to participate in future surveys). As a predictor of potential differences in these quality indicators, we use the advertisement design through which the respondent entered the survey.

Survey break-off rate is an important measure of data quality, because it indicates the respondents' overall motivation to respond to the whole questionnaire (Tangmanee & Niruttinanon, 2019). Previous research has shown how this rate can be influenced by survey design features, such as including a progress bar or announcing the survey length (Liu et al., 2016). We calculate the break-off rate as the ratio of the number of surveys started—defined as respondents proceeding past the welcome page, thereby agreeing to participate and consenting to the processing of personal data—to the total number of surveys. This is calculated separately for each advertisement design. In the regression analysis, P refers to the probability of having started but not completed the survey, as opposed to having completed the survey.

Next, we define the outcome of *speeding*. Providing answers quickly usually indicates that a respondent wants to finish the survey without giving enough thought to the questions to provide accurate answers (Zhang & Conrad, 2014). However, the interpretation of a short response time is not straightforward, as it could also indicate that the respondents have stable and crystallized opinions about certain topics, or that the survey design is efficient (Zhang & Conrad, 2014). Nevertheless, survey completion time is a commonly used data quality indicator that reflects possible general problems with the survey itself or the motivation of (some) respondents to answer the questions thoroughly. The survey completion times presented here are calculated based only on completed interviews—that is, surveys that reached the final page of the web survey, regardless of item

non-response. In the multivariate analysis, we transform the completion time into a binary variable, defining speeding as having a completion time in the fastest 10% of the sample distribution. This means that the completion time is less than 9.94 minutes for Germany and less than 10.60 minutes for the United States. *P* refers to the probability of speeding.

We analyze *non-differentiation* in the context of satisficing behavior. This behavior may result from a lack of motivation or response-ability (Gao et al., 2016; Roberts et al., 2019). We use a battery of eight items that measure attitudes toward immigration and estimate the number of inconsistent responses. The response scale ranged from fully agree to fully disagree on a five-point scale and included an option for no opinion, which was excluded from this analysis. For the immigration items, half of the statements were framed with positive attitudes toward immigration, and the other half were framed with negative attitudes toward immigration (see Appendix Table A3). We assume that an attentive respondent would tend to agree with half of the items and tend to disagree with the other half. In order to assess the consistency of response behavior, the rating scales of the items were re-coded to point in the same direction. In the following steps, the mean value was calculated for the group of positively and negatively framed statements. Next, the absolute difference between these means was calculated. If the responding behavior was consistent, the difference should be close to zero, while higher values should indicate inconsistent responding behavior. When respondents reached the maximum value of four, this means that they fully agreed with one framing and also fully agreed with items with the contradictory framing, indicating inconsistent response behavior. For the regression analysis, we categorize this outcome into two categories: low non-differentiation (0) is assigned for values below the median value of 0.5; while high non-differentiation (1) is assigned to all values above the threshold of the median value of 0.5. *P* refers to the probability of high non-differentiation.

Item non-response means that participants started to answer the questionnaire, but did not answer certain questions where a response would have been expected (Cehovin et al., 2023). Respondents may choose not to answer a particular question for many reasons. These include not knowing or remembering the answer, privacy concerns, or a lack of motivation. In this regard, research has shown that adding motivational statements after a question is left unanswered reduces item non-response in self-administered surveys (Al Baghal & Lynn, 2015). In our study, item non-response is defined as seeing a survey question but not responding to it. There were no compulsory questions in the survey. Providing non-substantive answers (e.g., "prefer not to say" or "other") does not count as non-response. Respondents who did not start the survey are excluded. The percentage of non-response to the survey questions was calculated as the percentage of missing responses (i.e., a question that was seen but not answered) divided by the number of expected responses, which is the sum of the number of

times a valid response was recorded and the number of times the question was seen and no response was recorded.

For the regression analysis, we categorize item non-response into two categories: zero for no item non-response and one if respondents did not respond to at least one item. *P* refers to the probability of item non-response.

Next, we consider *attentiveness*, which is measured by an attention check question in the form of an instructed response item (IRI) developed by Gummer et al., (2018). These items are included as part of a grid of questions in which one item asks respondents to select a particular response category. This assesses whether respondents have read the text of the particular item. Failure to provide the required response indicates inattention due to insufficient reading or understanding of the particular item (Gummer et al., 2018). In our survey, the IRI was administered in a list of six statements about politics and society (see Appendix Table A5 for the question text). Respondents were asked to indicate their opinion regarding these statements on a five-point scale, ranging from "strongly agree" to "strongly disagree." Item four of the six was not a political statement, but instructed respondents to choose a specific value on the response scale ("Please click 'rather disagree'."). From this, we construct our measure of attention by defining the attention check as "passed" (1) if the required category was selected, and as "failed" (0) if any other category was selected. Item non-response to this question was excluded before defining these categories. From this, we calculate the percentage of respondents who passed the attention check for each ad design. For the regression analysis, *P* refers to the probability of passing the attention check.

At the end of the survey, respondents were asked if they would like to provide their e-mail address so they could be contacted for future surveys. Loosveldt and Storms (2008) show that the *willingness to participate in future surveys* is influenced by the respondent's overall opinion of surveys. Willingness to participate in the future is increased when respondents perceive surveys as a useful tool for sharing their opinions. On the other hand, the likelihood of future participation is reduced when respondents perceive the investment of time and cognitive effort required as too high, or when there are concerns about data privacy (Loosveldt & Storms, 2008). Willingness to participate in future surveys is measured by the percentage of respondents who provided an e-mail address. This is calculated as the number of e-mail entries divided by the total; that is, the sum of the entries and the empty entries (i.e., the sum of respondents who saw the question and did not enter an e-mail in the open text field.). For the regression analysis, the binary outcome of providing an e-mail address (1) or not providing an e-mail address (0) is used. *P* refers to the probability of providing an e-mail address.

Regression Analysis

We use logistic regression to estimate the effects of advertisement design on each binary data quality outcome:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{ad_{imm-strong}} + \beta_2 x_{ad_{imm-loose}} + \beta_3 x_{ad_{clim-strong}} + \beta_4 x_{ad_{clim-loose}} + \beta_5 x_{sex} + \beta_6 x_{age_{40-59}} + \beta_7 x_{age_{60-64}} + \beta_8 x_{age_{65+}} + \beta_9 x_{income_{low}} + \beta_{10} x_{income_{median}} + \beta_{11} x_{income_{missing}} + \beta_{12} x_{device}$$

The independent variables in the models above correspond to the following categories:

- *Advertisement design* refers to the ad image through which a Facebook user reached our survey. These are: "immigration-strong" (imm-strong), "immigration-loose" (imm-loose), "climate-strong" (clim-strong), "climate-loose" (clim-loose), and the "neutral" design. We use the "neutral" design as the reference."
- *Sex* refers to the respondents' self-reported sex (i.e., male, female). We use "female" as the reference.
- *Age* represents the respondents' age, grouped into four age categories (i.e., 18–39, 40–59, 60–64, 65+). We use the "18-39" age group as the reference.
- *Income* refers to the monthly household income, with the categories "low" (≤ 25th percentile), "medium" (> 25th percentile and ≤ 50th percentile), "high" (> 50th percentile), and "missing." We use "high" as the reference.
- *Device* refers to the device that the respondents used to fill out the survey. It consists of the categories "mobile" (Android Smartphone/Tablet and iPad/iPhone) and "desktop." We use "desktop" as the reference.

For the regression analysis, the missing values for gender and age were removed. Full case analysis is considered more valid than using an imputation technique that would estimate socio-demographic information such as age and sex based on limited information in the survey. The analysis was conducted using R version 4.3.2. A complete list of the R packages used in the analysis can be found in the Appendix Section A4.

Results

Recruitment Results and Response Quality

In terms of campaign performance, in Germany, the design strongly related to immigration received the most link clicks (1,646) and the neutral design received the fewest clicks (859). In the United States, the design loosely related to climate change received the most clicks (2,778) and the neutral design received

the fewest clicks (1,273) (see Appendix Table A1). Using the impression and reach performance metrics provided by Meta (see Appendix Table A1), we calculate an indicator of the average number of times each ad was shown to a user. This shows that, on average, a user had a chance to see our ads between 1.3 to 1.4 times. Since we cannot assume that each user consciously saw the ads for each impression, this measure can be interpreted as an upper bound estimate and the vast majority of respondents most likely saw only one specific ad. In Germany about 890 euros were spent, with a cost per click of between 0.11 and 0.21 euros. Recruitment costs were higher in the United States, with 3,457 euros spent and a cost per click ranging from 0.25 to 0.55 euros.

Initially, 6,827 respondents in Germany and 12,596 in the United States reached our survey website. We had to exclude those cases with no information in the predictor variable of advertisement design (this removes 159 cases for Germany and 190 cases for the United States). The Facebook sample consisted of 6,668 respondents for Germany and 12,406 for the United States, including both started and not started surveys. In Germany, 4,170 respondents started the survey and 2,495 completed it. In the United States, 5,469 respondents started the survey and 2,520 completed it (see Appendix Table A2).

Among the respondents who started surveys in Germany, 31% were recruited through a design strongly related to immigration, while the smallest share (14%) were recruited through a design loosely related to climate. Among respondents who started surveys in the United States, 26% were recruited through the design strongly related to climate, while the smallest shares were recruited through the neutral and the design loosely related to immigration (each 16%) (see Appendix Table A1).

In both countries, the samples consist of about 50% men. Approximately 40% of all respondents are female, and 7–8% of the participants did not provide gender information. On average, the participants were 74 years old in the United States (range: 19–96, *SD* = 10) and were 60 years old in Germany (range: 18–99, *SD* = 11). However, 10% of respondents in Germany and 12% of respondents in the United States did not report their age (see Appendix Table A4).

Table 1 provides a descriptive overview of the quality indicators by country. Only started surveys are included in these results. The survey break-off rate is higher in the United States (54%) than in Germany (40%). On average, the respondents completed the survey in 16 minutes in Germany and 18 minutes in the United States. The rate of non-differentiation is higher in Germany (0.5) than in the United States (0.4). In both countries, there is about 3% item non-response to the survey questions. More participants passed the attention check question in the United States (77%) than in Germany (64%). Finally, the willingness to provide an e-mail address is higher in the United States (53%) than in Germany (42%) (see Table 1).

We also examine the changes in the quality indicators over the eight-day recruitment period by advertisement design. There are no systematic time

trends in the evolution of the quality indicators, suggesting that the algorithmic placement of the ads does not promote specific response quality types (see Appendix Figure A2).

Compared to the online panel respondents, the social media sample has a significantly higher break-off rate. However, online panel respondents were incentivized to complete the survey. Average completion time and item non-response rates are lower for the online panel than for the social media sample. The rate of passing the attention check is higher in the online panel. On the other hand, the rate of non-differentiation is higher in the online panel than in the social media sample. See Section A2 in the appendix for a description of the online panel sample.

Table 1 Descriptive statistics by country showing the survey quality indicators: break-off rate, mean completion time, non-differentiation, item non-response, and e-mail provision

Indicator	Germany	United States
Break-off rate (%)	40.17 [38.69, 41.66]	53.92 [52.60, 55.24]
Mean completion time (min)	16.11 [15.78, 16.43]	17.93 [17.57, 18.30]
Non-differentiation	0.51 [0.47, 0.54]	0.40 [0.37, 0.44]
Item non-response (%)	2.91 [2.85, 2.97]	3.19 [3.13, 3.26]
Attention check passed (%)	63.51 [61.57, 65.40]	76.73 [75.02, 78.35]
Provided e-mail address (%)	42.30 [40.47, 44.15]	52.85 [50.99, 54.70]

Notes: Values in brackets refer to the 95% confidence interval. Unweighted.

Advertisement Design Effects on Data Quality

This chapter presents the results of the binary logistic regression analysis used to assess the six quality indicators separately for Germany and the United States. Figure 2 shows the odds ratio estimates controlling for gender, age, income, and the device used to answer the survey, with the neutral design as the reference category.

We first present the results for Germany. We find a higher probability of leaving the survey (OR = 1.6, 95% CI [1.03, 2.42])[1] for the design classified as strongly related to immigration. In terms of speeding—that is, having a survey completion time in the top 10th percentile—we see that the design classified as strongly related to climate is associated with a lower likelihood of speeding than the neu-

[1] OR = odds ratio, CI = confidence interval.

tral ad design (OR = 0.6, 95% CI [0.41, 0.95]). The design classified as strongly related to immigration is associated with a lower chance of passing the attention check (OR = 0.6, 95% CI [0.43, 0.76]). Finally, respondents recruited by the design classified as having a strong topic relation (both immigration (OR = 0.5, 95% CI [0.39, 0.65]) and climate (OR = 0.6, 95% CI [0.45, 0.75])) are less likely to provide an e-mail address than respondents recruited by the neutral design. There is no design effect on non-differentiation and item non-response (see Figure 2).

For the United States, we see a correlation with a lower likelihood of speeding (OR = 0.5, 95% CI [0.32, 0.75] for climate and OR = 0.6, 95% CI [0.38, 0.88] for immigration) and a higher likelihood of non-differentiation (OR = 1.5, 95% CI [1.16, 2.00] for climate and OR = 1.4, 95% CI [1.08, 1.87] for immigration) for the ads classified as having a strong topic relation compared to the neutral ads. The design classified as strongly related to climate is also associated with a higher probability of item non-response (OR = 1.4, 95% CI [1.04, 1.87]) and a lower probability of passing the attention check (OR = 0.7, 95% CI [0.47, 0.94]). The results for panel availability show that respondents recruited through the design classified as having a strong immigration relation are associated with a lower probability (OR = 0.7, 95% CI [0.56, 0.96]) of providing their e-mail address. The break-off rate and the speeding behavior are not associated with any specific ad design (see Figure 2). It is worth noting that the thematically loosely associated design does not differ significantly from the neutral design in any of the final models for either country.

Full stepwise regression estimates can be found in Appendix Section A3. Looking at the stepwise regression estimates, we can see that for Germany, there is no association between ad design and non-differentiation or item non-response in any of the stepwise models (see Appendix Tables A13 and A15). Similarly, in the United States, survey break-off is not associated with any particular ad design in the separate model steps (see Appendix Table A10). In some cases, initial design effects can be explained by the sample composition. The lower chance of speeding for the designs classified as having a loose topic relation is explained by the age of the respondents (climate) and the device used to fill out the survey (immigration) (see Appendix Table A12). Participation through the design classified as strongly related to immigration is associated with a lower chance of passing the attention check question, this association is explained by the gender of the participants (see Appendix Table A18). For the design classified as loosely related to climate change, the correlation with a lower probability of panel availability is explained by age structure (see Appendix Table A20).

In summary, the designs classified as highly related to the survey topic are associated with lower response quality. In Germany, higher break-off rates, lower odds of speeding, passing the attention check, and panel availability are associated with ad designs classified as strongly related to the survey topic. In the United States, the designs classified as strongly topic-related are associated

with a lower likelihood of speeding, a higher likelihood of non-differentiation and item non-response, and lower rates of passing the attention check and providing an e-mail address.

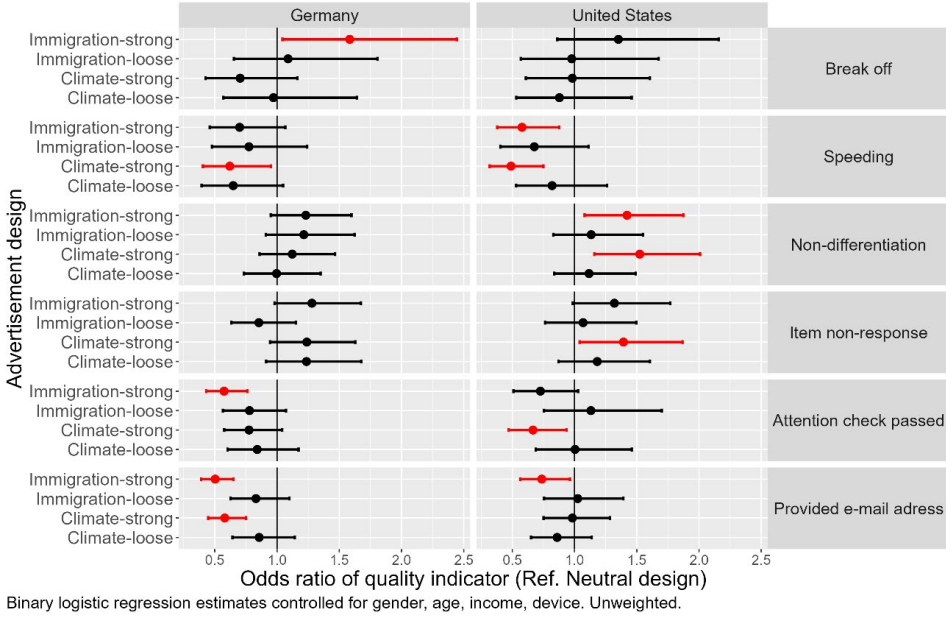

Figure 2 Stepwise binary logistic regression estimates for quality indicators, controlled for advertisement design, gender, age, income, and device for Germany and the United States. Unweighted. Odds ratios significantly different from one are marked in red.

Discussion

In this paper, we examined how advertisement design affects response quality in surveys recruited via social media. Using a study conducted via Facebook in Germany and the United States, we varied the design of the advertisement used for survey recruitment. We hypothesized that a more explicit display of the survey topic would result in systematically different response behavior compared to designs with fewer or no references to the survey topic. However, previous research was inconclusive about the direction of this effect.

We analyzed the impact of ad design on six response quality indicators: survey break-off rate, speeding, non-differentiation, item non-response, passing an attention check, and willingness to participate in a panel.

Consistent with previous findings, we observed that more respondents were recruited through advertisements with a prominent survey topic (i.e., immigration or climate change). However, after controlling for differences in sample composition by gender, age, income, and device, we found that these advertisements were associated with lower response quality. Specifically, ads classified as having a strong survey topic relation were associated with higher break-off rates, longer completion times, more non-differentiation, and more item nonresponse, as well as with lower rates of passing attention checks and willingness to participate in future surveys. We uncovered differences in ad design effects across countries: while ad design had no effect on break-off rates in the United States, it did in Germany; and while ad design had no effect on item nonresponse and non-differentiation in Germany, it did in the United States. We also found that the ads classified as less explicit (loose association) did not differ significantly from the neutral designs in terms of response quality after adjusting for sample composition. While these ads had higher reach and link clicks than neutral ads, they did not lead to higher start rates.

We considered two theoretical perspectives on how ad design might affect response quality. One perspective suggested that a more explicit display of the survey topic would attract highly thematically motivated respondents, thereby improving response quality. The other perspective posited that neutral ads would attract highly intrinsically motivated participants, leading to higher response quality. The results support the second argument, showing that neutral ad designs were associated with higher response quality than ads classified as having strong thematic references. This suggests that higher thematic motivation may lead to the recruitment of respondents with lower levels of response commitment, resulting in more inconsistent response behavior and lower overall response quality. This, in turn, supports our assumption that a more explicit display of the survey topic would lead to systematically different response styles than those expected from respondents recruited through ads with a less salient display of the survey topic. Finally, we want to address two emerging challenges for survey recruitment via social media advertisements: the rise of large language models and AI tools. One emerging challenge for data quality in social media-recruited surveys is the increasing use of large language models and chatbots to fabricate survey responses. These tools can undermine the authenticity and reliability of survey data, introducing new forms of bias and error. Depending on the complexity of the models, conventional survey quality indicators may not be able to distinguish between fabricated and genuine survey responses (Höhne et al. 2024). Future studies should proactively address this issue, as it is critical to maintaining the integrity of web survey-based research.

The use of generative AI tools to create ad images and text in social media recruitment campaigns, may, on the one hand, help to streamline the creative process, allowing for the rapid creation of visually engaging and personalized

content, increasing ad reach or engagement. However, it also raises issues of authenticity and audience trust. For example, overly polished or artificial-looking ads may increase user skepticism or reduce perceived credibility, which can impact survey participation rates. In addition, AI tools are trained on existing datasets, which can (re)introduce unintentional biases into ad images or messaging, which could affect the inclusivity and representativeness of survey samples. Future research is needed to explore this systematically. For example, A/B testing could be used to compare AI-generated ads with traditional ads in terms of response rates, participant demographics, and data quality.

Conclusion

Our study shows that advertisement design significantly affects response quality in social media-recruited surveys, with effects varying across different quality indicators and countries. Ads with higher topic salience tend to attract more clicks at a lower cost, but they often result in poorer response quality, including inconsistent responses and higher non-response rates. Conversely, neutral ads tend to yield higher response quality, making them more suitable for general research purposes. The findings have important implications for researchers planning future survey recruitment ad campaigns using Facebook. Specifically, there appears to be a trade-off between the level of attention generated by ads focused on prominent issues such as immigration and climate change (as indicated by higher reach and link clicks) and the quality of survey responses obtained.

While themed ads may initially lower recruitment cost and increase sample size, these benefits can be offset by a higher proportion of low-quality responses. The variation in design effects across countries also highlights the importance of considering country-specific contexts when designing a recruitment campaign that focuses on potentially polarizing social issues. Therefore, the specific objectives of the recruitment campaign should guide the choice of ad design.

However, our study also has several limitations that need to be acknowledged. Because we lacked a direct measure of the level of commitment evoked by the advertisements, we could only assume that higher topic salience correlates with higher thematic motivation.

The classification of images as loosely or strongly related to immigration or climate change is an individual and subjective interpretation and may vary from respondent to respondent. However, we base our image classification on our image selection survey, thus providing empirical support for our classification. Additionally, by excluding cases with missing age or sex information from the regression analysis, we may have underestimated the impact of ad design on response quality by removing the lowest-quality responses. Nevertheless, our models were robust to the inclusion of these missing values.

The generalizability of our findings is also limited by the exclusive use of Facebook for recruitment, which may not translate to other social media platforms such as Instagram, TikTok, or LinkedIn, each having different user demographics and engagement behaviors. In addition, an inherent limitation of studies using any social media platform is the underlying advertising algorithm, which remains a black box to researchers and may change over time, requiring frequent re-evaluation of any methodological finding to ensure robust results. Reliance on commercial platforms for data collection carries additional risks, as platform policies, access, and available features used for data collection, such as the business advertising manager for Meta, may change or be deprecated. This can limit data availability and affect the reproducibility of studies, as has been shown previously for studies using data obtained from social media platforms through API access (Davidson et al., 2023; Freelon, 2018).

In addition, the study focused on two topics that are heavily discussed and politically charged. While such topics are often the focus of social science research projects, our findings are limited, and might not be generalizable to other prominent but less controversial topics. Additionally, the study was conducted in Germany and the United States, where survey recruitment via ads is relatively common. It remains uncertain whether these findings will hold true in countries where this recruitment approach is newer, or where there is greater skepticism towards online ads or invitations.

Future research should address these limitations by exploring the impact of advertisement design across different social media platforms (such as Instagram, X, or TikTok) and a wider range of topics and contexts. By addressing these issues, future studies can build on our findings to further our understanding and optimize the use of social media advertisements to recruit survey participants.

Code and data availability

The data that support the findings of this study are available for scientific purposes upon request at: https://pub.uni-bielefeld.de/record/3002204

The code that was used for the analysis can be found at: osf.io/n76vu

Ethics approval and consent to participate

This study received ethics approval from the Ethics Council of the University of Bielefeld (Application Nr: 2022-209). Electronic informed consent was obtained from all participants who actively opted to participate in the online survey, enabling the collection, storage, and processing of their answers. All participants' data were treated anonymously. Participation was voluntary.

References

Al Baghal, T., & Lynn, P. (2015). Using motivational statements in web-instrument design to reduce item-missing rates in a mixed-mode context. *Public Opinion Quarterly, 79*(2), 568–579. https://doi.org/10.1093/poq/nfv023

Birkenmaier, L., Daikeler, J., Fröhling, L., Gummer, T., Lechner, C., Lux, V., Schwalbach, J., Silber, H., Weiß, B., Weller, K., Wolf, C., Abel, D., Breuer, J., Dietze, S., Dimitrov, D., Döring, H., Hebel, A., Hochman, O., Jünger, S., … Ziaja, S. (2024). *Defining and evaluating quality for the sciences: Position paper* (GESIS Papers, 2024/06). GESIS – Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.21241/ssoar.96764

Cehovin, G., Bosnjak, M., & Lozar Manfreda, K. (2023). Item nonresponse in web versus other survey modes: A systematic review and meta-analysis. *Social Science Computer Review, 41*(3), 926–945. https://doi.org/10.1177/08944393211056229

Choi, I., Milne, D. N., Glozier, N., Peters, D., Harvey, S. B., & Calvo, R. A. (2017). Using different Facebook advertisements to recruit men for an online mental health study: Engagement and selection bias. *Internet Interventions,* 8, 27–34. https://doi.org/10.1016/j.invent.2017.02.002

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., Van Der Linden, D., Roscoe, J. F., Ayravainen, L., & Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour, 7*(12), 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

De Man, J., Campbell, L., Tabana, H., & Wouters, E. (2021). The pandemic of online research in times of COVID-19. *BMJ Open, 11*(2), e043866. https://doi.org/10.1136/bmjopen-2020-043866

Donzowa, J., Perrotta, D., & Zagheni, E. (2023). *Assessing self-selection biases in online surveys: Evidence from the COVID-19 Health Behavior Survey* (MPIDR Working Paper WP-2023-047). Max Planck Institute for Demographic Research. https://doi.org/10.4054/MPIDR-WP-2023-047

Freelon, D. (2018). Computational research in the post-API age. *Political Communication, 35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online Survey. *Journal of Official Statistics, 22*(2), 313–328. https://www.proquest.com/docview/1266792615?pq-origsite=gscholar&fromopenview=true&sourcetype=Scholarly%20Journals

Gao, Z., House, L., & Bi, X. (2016). Impact of satisficing behavior in online surveys on consumer preference and welfare estimates. *Food Policy, 64*, 26–36. https://doi.org/10.1016/j.foodpol.2016.09.001

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *The Public Opinion Quarterly, 56*(4), 475–495. http://www.jstor.org/stable/2749203

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., Zagheni, E., Flores, R. D., Ventura, I., & Weber, I. (2022). Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(Supplement_2), S343–S363. https://doi.org/10.1111/rssa.12948

Gummer, T., Roßmann, J., & Silber, H. (2018). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research, 50*(1), 238-264. https://doi.org/10.1177/0049124118769083

Haer, R., & Meidert, N. (2013). Does the first impression count? Examining the effect of the welcome screen design on the response rate. *Survey Methodology, 39*(2), 419–434. http://www.statcan.gc.ca/pub/12-001-x/2013002/article/11885-eng.htm

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in a high internet-coverage population: Differences in response quality. *Public Opinion Quarterly, 72*(5), 836–846. https://www.jstor.org/stable/25548047

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*(1), 79–125. https://doi.org/10.1086/346010

Höhne, J. K., Claassen, J., Shahania, S., & Broneske, D. (2024). Bots in web survey interviews: A showcase. *International Journal of Market Research, 67*(1), 3–12. https://doi.org/10.1177/14707853241297009

Iannelli, L., Giglietto, F., Rossi, L., & Zurovac, E. (2020). Facebook digital traces for survey research: Assessing the efficiency and effectiveness of a Facebook ad–based procedure for recruiting online survey respondents in niche and difficult-to-reach populations. *Social Science Computer Review, 38*(4), 462–476. https://doi.org/10.1177/0894439318816638

Kaminska, O., McCutcheon, A. L., & Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly, 74*(5), 956–984. https://doi.org/10.1093/poq/nfq062

Kaplowitz, M. D., Lupi, F., Couper, M. P., & Thorp, L. (2012). The effect of invitation design on web survey response rates. *Social Science Computer Review, 30*(3), 339–349. https://doi.org/10.1177/0894439311419084

Keusch, F. (2013). The role of topic interest and topic salience in online panel web surveys. *International Journal of Market Research, 55*(1), 59–80. https://doi.org/10.2501/IJMR-2013-007

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Kühne, S., & Zindel, Z. (2020). Using Facebook and Instagram to recruit web survey participants: A step-by-step guide and application. *Survey Methods: Insights from the Field, Special Issue: 'Advancements in Online and Mobile Survey Methods'*. https://doi.org/10.13094/SMIF-2020-00017

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2021). Social media, web, and panel surveys: Using non-probability samples in social and policy research. *Policy & Internet, 13*(1), 134–155. https://doi.org/10.1002/poi3.238

Liu, M., Kuriakose, N., Cohen, J., & Cho, S. (2016). Impact of web survey invitation design on survey participation, respondents, and survey responses. *Social Science Computer Review, 34*(5), 631–644. https://doi.org/10.1177/0894439315605606

Loosveldt, G., & Storms, V. (2008). Measuring public opinions about surveys. *International Journal of Public Opinion Research, 20*(1), 74–89. https://doi.org/10.1093/ijpor/edn006

Mavletova, A., Deviatko, I., & Maloshonok, N. (2014). Invitation design elements in web surveys – Can one ignore interactions? *Bulletin of Sociological Methodology/Bulletin de Methodologie Sociologique, 123*(1), 68–79. https://doi.org/10.1177/0759106314531883

Neundorf, A., & Öztürk, A. (2022). *Advertising online surveys on social media: How your advertisements affect your samples*. OSF. https://doi.org/10.31219/osf.io/84h3t

Neundorf, A., & Öztürk, A. (2023). How to improve representativeness and cost- effectiveness in samples recruited through meta: A comparison of advertisement tools. *PLOS ONE, 18*(2), e0281243. https://doi.org/10.1371/journal.pone.0281243

Pötzschke, S., & Braun, M. (2017). Migrant sampling using Facebook advertisements: A case study of polish migrants in four European countries. *Social Science Computer Review, 35*(5), 633–653. https://doi.org/10.1177/0894439316666262

Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly, 83*(3), 598–626. https://doi.org/10.1093/poq/nfz035

Stern, M. J., Fordyce, E., Carpenter, R., Viox, M. H., Michaels, S., Harper, C., Johns, M. M., & Dunville, R. (2022). Evaluating the data quality of a national sample of young sexual and gender minorities recruited using social media: The influence of different design formats. *Social Science Computer Review, 40*(3), 663–677. https://doi.org/10.1177/0894439320928240

Tangmanee, C., & Niruttinanon, P. (2019). Web survey's completion rates: Effects of forced responses, question display styles, and subjects' attitude. *International Journal of Research in Business and Social Science (2147-4478), 8*(1), 20–29. https://doi.org/10.20525/ijrbs.v8i1.183

Wenemark, M., Persson, A., Brage, H. N., Svensson, T., & Kristenson, M. (2011). Applying motivation theory to achieve increased response rates, respondent satisfaction and data quality. *Journal of Official Statistics, 27*(2), 393–414. https://www.proquest.com/docview/2821376428?pq-origsite=gscholar&fromopenview=true&sourcetype=Scholarly%20Journals

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods, 8*(2), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453

Zillmann, D., Schmitz, A., Skopek, J., & Blossfeld, H.-P. (2014). Survey topic and unit nonresponse. *Quality & Quantity, 48*(4), 2069–2088. https://doi.org/10.1007/s11135-013-9880-y

Zindel, Z. (2023). Social media recruitment in online survey research: A systematic literature review. *methods, data, analyses, 17*(2), 207–248. https://doi.org/10.12758/mda.2022.15

Appendix

A1 Descriptive Statistics and Survey Items

Figure A1 Advertisement images used in the Facebook ads to recruit respondents for the survey with varying associations to the survey topics of immigration and climate change. None of the images used in our study were generated (or partially generated/altered) by AI.

Table A1 Campaign performance and number of started surveys in Germany and the United States

Advertisement	Impression	Reach	Unique link click	Impression to reach ratio	Cost per unique link click (in €)	Started survey	
Germany						Count	%
Immigration-strong	25,008	17,976	1,646	1.39	0.11	1275	30.6
Immigration-loose	23,371	18,728	1,116	1.25	0.16	654	15.7
Climate-strong	24,804	18,568	1,176	1.34	0.15	929	22.2
Climate-loose	28,067	22,088	1,384	1.27	0.13	578	13.9
Neutral	19,929	14,864	859	1.34	0.21	737	17.7
United States							
Immigration-strong	43,191	32,160	2,567	1.34	0.27	1345	24.6
Immigration-loose	52,410	38,592	2,684	1.36	0.26	867	15.9
Climate-strong	50,128	36,759	2,447	1.36	0.28	1428	26.1
Climate-loose	65,716	50,608	2,778	1.30	0.25	954	17.4
Neutral	27,079	18,888	1,273	1.43	0.55	875	16.0

Table A2 Data cleaning of the survey

Design	Immigration-strong	Immigration-loose	Climate-strong	Climate-loose	Neutral	Missing	Total
Germany							
Survey landing page hits	1,837	1,168	1,245	1,437	981	159	6,827
Exclusion of cases without ad information	1,837	1,168	1,245	1,437	981	0	6,668
Started surveys	1,275	654	926	578	737	0	4,170
Completed surveys	645	392	625	359	474	0	2,495
United States							
Survey landing page hits	2,757	2,819	2,547	2,763	1,520	190	12,596
Exclusion of cases without ad information	2,757	2,819	2,547	2,763	1,520	0	12,406
Started surveys	1,345	867	1,428	954	875	0	5,469
Completed surveys	559	354	716	479	412	0	2,520

Table A3 Question text measuring attitudes toward immigration in the survey for Germany and the United States used for calculation of inconsistent responses

Positive framing	
1.	Legal immigrants to America/Germany who are not citizens should have the same rights as American citizens.
2.	Immigrants are generally good for America's/Germany's economy.
3.	Legal immigrants should have equal access to public education as American citizens.
4.	Immigrants improve American/German society by bringing new ideas and cultures.
Negative framing	
1.	American/German culture is generally undermined by immigrants.
2.	Immigrants increase crime rates.
3.	America/Germany should take stronger measures to exclude illegal immigrants.
4.	Immigrants take jobs away from people who were born in America/Germany.

Table A4 Gender and age composition of the survey

(a) Gender composition

Country	Female	Male	Missing	Total
Germany	1533 (41%)	1974 (52%)	271 (7%)	3778 (100%)
United States	1881 (42%)	2208 (50%)	361 (8%)	4450 (100%)

(b) Age composition

Country	*M* age	*SD*	Minimum	Maximum	Missing
Germany	60	11	18	99	10%
United States	74	10	19	96	12%

Notes: Unweighted.

Table A5 Attention check question implemented in the survey

	Questionnaire text
Question text	Here are some common statements on politics and society. Please state whether you agree or disagree.
Statements	a Politicians care about what ordinary people think. b People like me do not have any influence on the government. c Politics is so complicated people like me are not able to understand what is going on. d Please click "rather disagree." e Citizens lack possibilities to influence politics. f In a democracy it is the duty of all citizens to vote regularly in elections.
Response scale	1 strongly disagree 2 rather disagree 3 neither agree nor disagree 4 rather agree 5 strongly agree

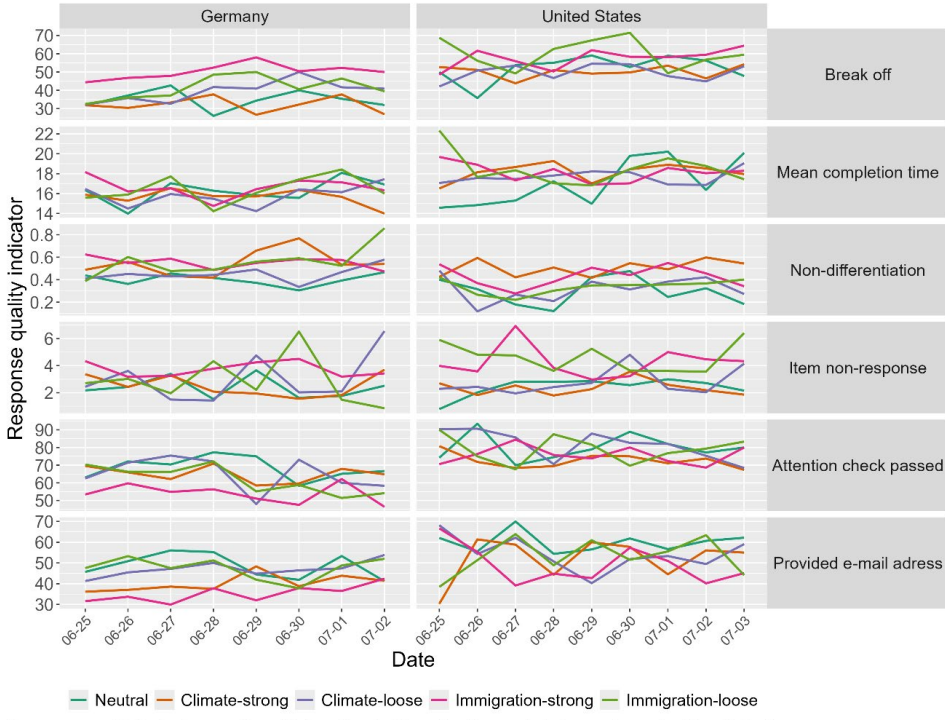

Figure A2 Descriptive statistics for the response quality indicator over the recruitment period by advertisement design for Germany and the United States. Unweighted.

A2 Comparison of Quality Indicators with the Online Panel

In addition to the social media recruitment efforts, a commercial panel company was asked to recruit a representative sample in both countries using the same survey questions. In this baseline sample, no promotional images were employed, and no information regarding the survey topics was provided (thus corresponding to the neutral design in the social media recruitment). For the reference sample, we received 1,555 surveys for Germany and 1,576 surveys for the United States from the company that fulfilled the inclusion criteria. As expected from an online panel, almost everyone completed the survey, with only 56 respondents in Germany and 55 respondents in the United States not completing the survey (see Table A7).

A3 Regression Results

Table A6 Number of started and completed surveys from the online panel

	Germany	United States
Started surveys	1,555	1,576
Completed surveys	1,499	1,521

A quota sampling approach was employed that takes the respective population composition of Germany and the United States into account. Thus, the gender composition of the sample is balanced. The mean age of the sample is 50 years for Germany and 48 years for the United States (see Table A7). Looking at the quality indicators, we see a very low break-off rate of less than 4% in both countries. The average completion time is about 11 minutes in both countries. Non-differentiation is higher in Germany, at 0.6, than in the United States, at 0.5. The item non-response rate is 0.3 in the United States and 0.6 in Germany. The percentage of respondents passing the attention check question is quite high, at 94% in Germany and 89% in the United States (see Table A8).

Table A7 Gender and age composition of the online panel

(a) Gender composition

Country	Female	Male	Total
Germany	790 (51%)	765 (49%)	1,555 (100%)
United States	793 (50%)	783 (50%)	1,576 (100%)

(b) Age composition

Country	*M* age	*SD*
Germany	50	17
United States	48	17

Notes: Unweighted. There are no missing values as this question was mandatory.

Table A8 Descriptive statistics by country for the online panel showing the survey quality indicators: break-off rate, mean completion time, non-differentiation, item non-response, attention check

	Germany		United States	
Break-off rate (%)	3.60	[2.78, 4.65]	3.49	[2.69, 4.51]
Mean completion time (min)	11.75	[11.27, 12.23]	11.34	[10.83, 11.84]
Non-differentiation	0.63	[0.58, 0.67]	0.46	[0.41, 0.50]
Item non-response (%)	0.58	[0.54, 0.63]	0.34	[0.31, 0.38]
Attention check passed (%)	94.48	[93.21, 95.53]	88.77	[87.09, 90.26]

Notes: Values in brackets refer to the 95% confidence interval. Unweighted.

A3 Regression Results

Table A9 Stepwise regression results for Germany, outcome: break-off

Stepwise regression results for Germany, outcome: break-off					
	Odds ratio estimate for binary logistic regression, Germany				
	Outcome: break-off (ref. completion)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	0.702 [0.432, 1.140]	0.751 [0.459, 1.227]	0.744 [0.453, 1.221]	0.747 [0.454, 1.231]	0.704 [0.426, 1.162]
Design: climate-loose	0.991 [0.589, 1.666]	1.003 [0.597, 1.687]	0.998 [0.592, 1.682]	1.037 [0.613, 1.754]	0.971 [0.573, 1.647]
Design: immigration-strong	1.665* [1.101, 2.519]	1.742** [1.148, 2.643]	1.747** [1.149, 2.656]	1.676* [1.099, 2.557]	1.584* [1.036, 2.422]
Design: immigration-loose	1.072 [0.652, 1.761]	1.131 [0.686, 1.866]	1.112 [0.674, 1.836]	1.131 [0.683, 1.874]	1.089 [0.656, 1.807]
Sex: male		0.773† [0.578, 1.034]	0.786 [0.587, 1.052]	0.877 [0.651, 1.180]	0.911 [0.676, 1.228]
Age: 40–59			1.595 [0.678, 3.752]	1.575 [0.666, 3.727]	1.585 [0.669, 3.754]
Age: 60–64			2.007 [0.834, 4.830]	1.875 [0.774, 4.539]	1.873 [0.773, 4.540]
Age: 65+			1.534 [0.645, 3.644]	1.365 [0.570, 3.267]	1.362 [0.568, 3.265]
Income: low				1.567* [1.041, 2.359]	1.617* [1.073, 2.437]
Income: medium				0.936 [0.601, 1.459]	0.939 [0.602, 1.463]
Income: missing				2.710** [1.857, 3.954]	2.766** [1.894, 4.040]
Device: mobile					2.120* [1.094, 4.109]
Constant	0.084** [0.060, 0.118]	0.093** [0.065, 0.133]	0.057** [0.023, 0.140]	0.041** [0.016, 0.105]	0.021** [0.007, 0.064]
Observations	2,455	2,455	2,455	2,455	2,455
Log likelihood	-702.957	-701.457	-699.641	-682.934	-679.937
AIC	1,415.915	1,414.914	1,417.282	1,389.868	1,385.873

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A10 Stepwise regression results for the United States, outcome: break-off

Stepwise regression results for the United States, outcome: break-off					
	Odds ratio estimate for binary logistic regression, United States				
	Outcome: break-off (ref. completion)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	0.992 [0.629, 1.563]	1.171 [0.732, 1.874]	1.147 [0.714, 1.843]	1.088 [0.675, 1.754]	0.982 [0.606, 1.592]
Design: climate-loose	1.017 [0.624, 1.660]	1.033 [0.633, 1.686]	0.986 [0.601, 1.616]	0.975 [0.593, 1.605]	0.878 [0.531, 1.452]
Design: immigration-strong	1.455† [0.931, 2.274]	1.518† [0.970, 2.377]	1.472† [0.938, 2.311]	1.396 [0.886, 2.201]	1.352 [0.856, 2.137]
Design: immigration-loose	1.098 [0.654, 1.846]	1.214 [0.719, 2.051]	1.168 [0.689, 1.979]	1.147 [0.674, 1.953]	0.977 [0.570, 1.674]
Sex: male		0.654** [0.488, 0.877]	0.660** [0.491, 0.886]	0.768† [0.567, 1.039]	0.833 [0.613, 1.132]
Age: 40–59			2.633 [0.327, 21.178]	2.827 [0.349, 22.878]	2.914 [0.360, 23.621]
Age: 60–64			1.951 [0.240, 15.866]	1.964 [0.240, 16.048]	2.049 [0.250, 16.782]
Age: 65+			3.043 [0.411, 22.513]	2.761 [0.371, 20.532]	2.789 [0.374, 20.778]
Income: low				0.931 [0.531, 1.631]	0.926 [0.528, 1.624]
Income: medium				0.917 [0.563, 1.492]	0.902 [0.554, 1.470]
Income: missing				2.459** [1.552, 3.895]	2.372** [1.496, 3.762]
Device: mobile					2.237** [1.392, 3.593]
Constant	0.082** [0.057, 0.118]	0.096** [0.066, 0.139]	0.034** [0.005, 0.249]	0.026** [0.003, 0.201]	0.014** [0.002, 0.110]
Observations	2,571	2,571	2,571	2,571	2,571
Log likelihood	-739.125	-735.095	-733.259	-712.284	-705.729
AIC	1,488.250	1,482.190	1,484.517	1,448.569	1,437.458

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A11 Stepwise regression results for Germany, outcome: speeding

Stepwise regression results for Germany, outcome: speeding					
	Odds ratio estimate for binary logistic regression, Germany				
	Outcome: speeding (ref. not speeding)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	0.748 [0.500, 1.120]	0.775 [0.514, 1.168]	0.595* [0.390, 0.907]	0.592* [0.388, 0.902]	0.620* [0.405, 0.950]
Design: climate-loose	0.738 [0.459, 1.186]	0.744 [0.463, 1.196]	0.625† [0.385, 1.014]	0.618† [0.381, 1.004]	0.648† [0.398, 1.055]
Design: immigration-strong	0.766 [0.511, 1.146]	0.784 [0.522, 1.178]	0.673† [0.444, 1.021]	0.672† [0.443, 1.020]	0.700† [0.460, 1.065]
Design: immigration-loose	0.741 [0.467, 1.175]	0.764 [0.480, 1.216]	0.753 [0.468, 1.210]	0.749 [0.466, 1.205]	0.774 [0.481, 1.247]
Sex: male		0.873 [0.657, 1.161]	0.863 [0.645, 1.154]	0.840 [0.626, 1.126]	0.816 [0.607, 1.096]
Age: 40–59			0.563* [0.338, 0.937]	0.560* [0.336, 0.934]	0.558* [0.335, 0.931]
Age: 60–64			0.472** [0.270, 0.825]	0.482* [0.275, 0.844]	0.481* [0.275, 0.843]
Age: 65+			0.142** [0.077, 0.260]	0.149** [0.081, 0.274]	0.149** [0.081, 0.274]
Income: low				0.736 [0.490, 1.105]	0.718 [0.478, 1.080]
Income: medium				0.790 [0.547, 1.141]	0.788 [0.545, 1.138]
Income: missing				0.792 [0.517, 1.211]	0.778 [0.508, 1.192]
Device: mobile					0.664† [0.427, 1.033]
Constant	0.138** [0.103, 0.185]	0.146** [0.107, 0.200]	0.409** [0.232, 0.723]	0.476* [0.264, 0.861]	0.682 [0.337, 1.382]
Observations	2,247	2,247	2,247	2,247	2,247
Log likelihood	-725.305	-724.871	-692.755	-691.148	-689.606
AIC	1,460.610	1,461.742	1,403.511	1,406.296	1,405.213

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A12 Stepwise regression results for the United States, outcome: speeding

Stepwise regression results for the United States, outcome: speeding					
	Odds ratio estimate for binary logistic regression, United States				
	Outcome: speeding (ref. not speeding)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	0.372** [0.253, 0.548]	0.360** [0.241, 0.538]	0.424** [0.279, 0.644]	0.425** [0.279, 0.649]	0.489** [0.319, 0.751]
Design: climate-loose	0.514** [0.344, 0.766]	0.512** [0.343, 0.764]	0.712 [0.467, 1.086]	0.706 [0.462, 1.079]	0.820 [0.532, 1.262]
Design: immigration-strong	0.476** [0.322, 0.704]	0.471** [0.318, 0.698]	0.577** [0.382, 0.871]	0.578** [0.382, 0.876]	0.578* [0.380, 0.879]
Design: immigration-loose	0.420** [0.264, 0.669]	0.410** [0.256, 0.657]	0.520** [0.318, 0.850]	0.521** [0.318, 0.853]	0.677 [0.408, 1.122]
Sex: male		1.095 [0.826, 1.451]	0.978 [0.732, 1.307]	0.914 [0.680, 1.228]	0.799 [0.590, 1.081]
Age: 40–59			0.521 [0.239, 1.137]	0.501† [0.228, 1.102]	0.467† [0.210, 1.037]
Age: 60–64			0.404* [0.184, 0.887]	0.395* [0.179, 0.871]	0.352* [0.158, 0.786]
Age: 65+			0.105** [0.052, 0.215]	0.106** [0.052, 0.218]	0.098** [0.047, 0.203]
Income: low				0.617* [0.403, 0.945]	0.625* [0.407, 0.961]
Income: medium				0.613** [0.424, 0.886]	0.620* [0.427, 0.901]
Income: missing				0.418** [0.269, 0.651]	0.436** [0.279, 0.682]
Device: mobile					0.426** [0.312, 0.582]
Constant	0.219** [0.169, 0.284]	0.211** [0.159, 0.281]	1.248 [0.610, 2.555]	2.089† [0.955, 4.567]	4.078** [1.778, 9.354]
Observations	2,355	2,355	2,355	2,355	2,355
Log likelihood	-767.650	-767.450	-717.532	-709.904	-696.205
AIC	1,545.300	1,546.900	1,453.064	1,443.807	1,418.410

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A13 Stepwise regression results for Germany, outcome: non-differentiation

Stepwise regression results for Germany, outcome: non-differentiation					
	Odds ratio estimate for binary logistic regression, Germany				
	Outcome: non-differentiation (ref. no non-differentiation)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	1.171 [0.906, 1.514]	1.102 [0.849, 1.432]	1.127 [0.865, 1.469]	1.135 [0.871, 1.480]	1.122 [0.859, 1.465]
Design: climate-loose	1.000 [0.741, 1.350]	0.984 [0.728, 1.329]	0.998 [0.738, 1.350]	1.010 [0.746, 1.367]	0.996 [0.735, 1.351]
Design: immigration-strong	1.285† [0.996, 1.657]	1.233 [0.953, 1.594]	1.245† [0.962, 1.612]	1.244† [0.960, 1.612]	1.231 [0.949, 1.597]
Design: immigration-loose	1.263 [0.949, 1.679]	1.206 [0.904, 1.608]	1.217 [0.912, 1.623]	1.225 [0.918, 1.635]	1.215 [0.910, 1.623]
Sex: male		1.256* [1.053, 1.497]	1.246* [1.045, 1.486]	1.289** [1.078, 1.541]	1.300** [1.086, 1.555]
Age: 40–59			0.843 [0.557, 1.278]	0.845 [0.557, 1.282]	0.846 [0.558, 1.284]
Age: 60–64			0.788 [0.509, 1.221]	0.770 [0.496, 1.196]	0.770 [0.496, 1.196]
Age: 65+			0.930 [0.610, 1.419]	0.895 [0.586, 1.369]	0.895 [0.585, 1.369]
Income: low				1.284* [1.013, 1.628]	1.294* [1.020, 1.641]
Income: medium				1.149 [0.920, 1.435]	1.150 [0.920, 1.436]
Income: missing				1.269† [0.989, 1.629]	1.274† [0.993, 1.636]
Device: mobile					1.138 [0.844, 1.534]
Constant	0.676** [0.557, 0.821]	0.612** [0.496, 0.754]	0.701 [0.449, 1.094]	0.614* [0.388, 0.974]	0.547* [0.321, 0.933]
Observations	2,201	2,201	2,201	2,201	2,201
Log likelihood	-1,505.645	-1,502.405	-1,501.147	-1,498.229	-1,497.867
AIC	3,021.289	3,016.811	3,020.293	3,020.457	3,021.734

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A14 Stepwise regression results for the United States, outcome: non-differentiation

Stepwise regression results for the United States, outcome: non-differentiation					
	Odds ratio estimate for binary logistic regression, United States				
	Outcome: non-differentiation (ref. no non-differentiation)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	1.738** [1.341, 2.253]	1.618** [1.238, 2.115]	1.608** [1.227, 2.108]	1.627** [1.240, 2.134]	1.524** [1.158, 2.006]
Design: climate-loose	1.210 [0.913, 1.603]	1.201 [0.905, 1.592]	1.175 [0.883, 1.563]	1.188 [0.893, 1.581]	1.117 [0.836, 1.491]
Design: immigration-strong	1.481** [1.130, 1.941]	1.452** [1.107, 1.905]	1.430* [1.087, 1.879]	1.440** [1.095, 1.895]	1.421* [1.080, 1.872]
Design: immigration-loose	1.318† [0.976, 1.779]	1.258 [0.929, 1.704]	1.232 [0.907, 1.673]	1.247 [0.918, 1.694]	1.133 [0.829, 1.549]
Sex: male		1.204* [1.013, 1.432]	1.208* [1.015, 1.437]	1.237* [1.036, 1.477]	1.297** [1.084, 1.553]
Age: 40–59			1.713 [0.735, 3.992]	1.758 [0.755, 4.096]	1.801 [0.771, 4.204]
Age: 60–64			1.260 [0.543, 2.925]	1.271 [0.548, 2.948]	1.322 [0.569, 3.072]
Age: 65+			1.667 [0.760, 3.657]	1.702 [0.776, 3.733]	1.726 [0.785, 3.794]
Income: low				1.252 [0.944, 1.660]	1.241 [0.936, 1.647]
Income: medium				1.035 [0.811, 1.320]	1.026 [0.803, 1.309]
Income: missing				1.117 [0.861, 1.449]	1.094 [0.842, 1.420]
Device: mobile					1.451** [1.162, 1.813]
Constant	0.500** [0.405, 0.617]	0.464** [0.372, 0.580]	0.288** [0.131, 0.634]	0.254** [0.113, 0.572]	0.190** [0.083, 0.437]
Observations	2,421	2,421	2,421	2,421	2,421
Log likelihood	-1,626.546	-1,624.329	-1,622.075	-1,620.441	-1,614.937
AIC	3,263.092	3,260.658	3,262.151	3,264.881	3,255.874

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A15 Stepwise regression results for Germany, outcome: item non-response

Stepwise regression results for Germany, outcome: item non-response					
	Odds ratio estimate for binary logistic regression, Germany				
	Outcome: item non-response (ref. no item non-response)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	1.017 [0.799, 1.295]	1.088 [0.851, 1.391]	1.194 [0.929, 1.534]	1.236 [0.943, 1.622]	1.239 [0.944, 1.628]
Design: climate-loose	1.071 [0.812, 1.413]	1.085 [0.822, 1.433]	1.146 [0.865, 1.517]	1.234 [0.911, 1.670]	1.237 [0.912, 1.676]
Design: immigration-strong	1.223† [0.963, 1.553]	1.279* [1.005, 1.629]	1.356* [1.062, 1.731]	1.277† [0.979, 1.667]	1.280† [0.980, 1.672]
Design: immigration-loose	0.822 [0.627, 1.078]	0.868 [0.660, 1.140]	0.856 [0.650, 1.126]	0.853 [0.632, 1.150]	0.854 [0.633, 1.152]
Sex: male		0.773** [0.656, 0.910]	0.783** [0.664, 0.924]	0.880 [0.734, 1.054]	0.878 [0.731, 1.054]
Age: 40–59			1.960** [1.295, 2.967]	1.931** [1.231, 3.028]	1.931** [1.231, 3.028]
Age: 60–64			2.538** [1.643, 3.920]	2.455** [1.532, 3.935]	2.455** [1.532, 3.935]
Age: 65+			2.850** [1.869, 4.344]	2.685** [1.697, 4.247]	2.685** [1.698, 4.248]
Income: low				1.026 [0.818, 1.288]	1.025 [0.816, 1.287]
Income: medium				0.979 [0.789, 1.213]	0.978 [0.789, 1.213]
Income: missing				16.799** [11.481, 24.580]	16.781** [11.467, 24.558]
Device: mobile					0.972 [0.713, 1.324]
Constant	0.983 [0.819, 1.179]	1.098 [0.903, 1.335]	0.460** [0.295, 0.716]	0.309** [0.188, 0.506]	0.317** [0.180, 0.559]
Observations	2,455	2,455	2,455	2,455	2,455
Log likelihood	-1,696.657	-1,691.898	-1,674.754	-1,469.936	-1,469.920
AIC	3,403.314	3,395.795	3,367.508	2,963.873	2,965.840

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A16 Stepwise regression results for the United States, outcome: item non-response

Stepwise regression results for the United States, outcome: item non-response					
	Odds ratio estimate for binary logistic regression, United States				
	Outcome: item non-response (ref. no item non-response)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	1.233† [0.967, 1.571]	1.474** [1.144, 1.898]	1.408** [1.090, 1.819]	1.365* [1.023, 1.822]	1.393* [1.040, 1.866]
Design: climate-loose	1.210 [0.931, 1.573]	1.233 [0.947, 1.606]	1.147 [0.877, 1.499]	1.161 [0.858, 1.571]	1.182 [0.871, 1.604]
Design: immigration-strong	1.388* [1.078, 1.787]	1.459** [1.130, 1.883]	1.387* [1.072, 1.795]	1.315† [0.982, 1.761]	1.320† [0.986, 1.767]
Design: immigration-loose	1.009 [0.759, 1.341]	1.125 [0.843, 1.502]	1.061 [0.792, 1.420]	1.038 [0.746, 1.444]	1.069 [0.764, 1.496]
Sex: male		0.634** [0.537, 0.748]	0.645** [0.546, 0.762]	0.849† [0.702, 1.028]	0.837† [0.689, 1.015]
Age: 40–59			1.257 [0.557, 2.835]	1.618 [0.652, 4.015]	1.606 [0.647, 3.990]
Age: 60–64			1.512 [0.678, 3.375]	1.734 [0.705, 4.262]	1.716 [0.697, 4.226]
Age: 65+			2.118* [1.004, 4.471]	2.101† [0.907, 4.864]	2.093† [0.903, 4.854]
Income: low				1.325† [0.991, 1.773]	1.326† [0.991, 1.773]
Income: medium				1.133 [0.877, 1.462]	1.135 [0.879, 1.465]
Income: missing				12.481** [9.283, 16.781]	12.562** [9.339, 16.898]
Device: mobile					0.901 [0.717, 1.132]
Constant	0.725** [0.598, 0.880]	0.865 [0.705, 1.062]	0.451* [0.213, 0.952]	0.183** [0.077, 0.437]	0.199** [0.082, 0.483]
Observations	2,571	2,571	2,571	2,571	2,571
Log likelihood	-1,770.216	-1,755.555	-1,747.619	-1,462.862	-1,462.464
AIC	3,550.431	3,523.110	3,513.238	2,949.725	2,950.927

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A17 Stepwise regression results for Germany, outcome: attention check passed

Stepwise regression results for Germany, outcome: attention check passed					
	Odds ratio estimate for binary logistic regression, Germany				
	Outcome: attention check passed (ref. attention check failed)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	0.826 [0.625, 1.093]	0.847 [0.637, 1.125]	0.757† [0.566, 1.012]	0.737* [0.549, 0.988]	0.775† [0.576, 1.042]
Design: climate-loose	0.881 [0.638, 1.216]	0.886 [0.642, 1.224]	0.836 [0.603, 1.159]	0.791 [0.569, 1.101]	0.841 [0.603, 1.172]
Design: immigration-strong	0.587** [0.447, 0.772]	0.597** [0.453, 0.787]	0.553** [0.418, 0.731]	0.546** [0.412, 0.725]	0.574** [0.432, 0.763]
Design: immigration-loose	0.761† [0.558, 1.038]	0.775 [0.566, 1.060]	0.777 [0.567, 1.065]	0.758† [0.551, 1.042]	0.778 [0.565, 1.072]
Sex: male		0.914 [0.757, 1.103]	0.917 [0.758, 1.108]	0.845† [0.697, 1.026]	0.817* [0.673, 0.993]
Age: 40–59			0.874 [0.558, 1.369]	0.859 [0.545, 1.352]	0.854 [0.542, 1.346]
Age: 60–64			0.518** [0.324, 0.828]	0.537* [0.334, 0.864]	0.534** [0.332, 0.859]
Age: 65+			0.524** [0.332, 0.826]	0.573* [0.362, 0.909]	0.575* [0.362, 0.912]
Income: low				0.543** [0.422, 0.699]	0.526** [0.408, 0.677]
Income: medium				0.635** [0.500, 0.808]	0.632** [0.497, 0.804]
Income: missing				0.493** [0.374, 0.650]	0.485** [0.368, 0.640]
Device: mobile					0.534** [0.376, 0.760]
Constant	2.306** [1.863, 2.854]	2.399** [1.908, 3.017]	3.800** [2.333, 6.190]	5.646** [3.380, 9.433]	9.981** [5.430, 18.347]
Observations	2,081	2,081	2,081	2,081	2,081
Log likelihood	-1,348.735	-1,348.296	-1,332.194	-1,313.826	-1,307.277
AIC	2,707.469	2,708.591	2,682.387	2,651.652	2,640.555

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A18 Stepwise regression results for the United States, outcome: attention check passed

Stepwise regression results for the United States, outcome: attention check passed					
	Odds ratio estimate for binary logistic regression, United States				
	Outcome: attention check passed (ref. attention check failed)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	0.559** [0.404, 0.772]	0.587** [0.421, 0.819]	0.584** [0.417, 0.818]	0.586** [0.418, 0.821]	0.666* [0.473, 0.939]
Design: climate-loose	0.888 [0.617, 1.278]	0.893 [0.620, 1.286]	0.895 [0.619, 1.295]	0.891 [0.615, 1.290]	1.004 [0.690, 1.461]
Design: immigration-strong	0.705* [0.501, 0.993]	0.718† [0.509, 1.012]	0.718† [0.507, 1.015]	0.719† [0.508, 1.018]	0.726† [0.511, 1.031]
Design: immigration-loose	0.896 [0.606, 1.323]	0.927 [0.625, 1.375]	0.927 [0.623, 1.380]	0.915 [0.614, 1.363]	1.132 [0.754, 1.698]
Sex: male		0.874 [0.701, 1.091]	0.873 [0.699, 1.089]	0.834 [0.666, 1.045]	0.745* [0.592, 0.937]
Age: 40–59			1.339 [0.543, 3.300]	1.292 [0.523, 3.193]	1.212 [0.487, 3.019]
Age: 60–64			1.421 [0.578, 3.494]	1.416 [0.575, 3.488]	1.303 [0.525, 3.237]
Age: 65+			1.195 [0.528, 2.704]	1.201 [0.529, 2.724]	1.142 [0.500, 2.611]
Income: low				0.918 [0.647, 1.303]	0.931 [0.654, 1.327]
Income: medium				1.060 [0.782, 1.436]	1.086 [0.799, 1.475]
Income: missing				0.758† [0.548, 1.049]	0.794 [0.572, 1.102]
Device: mobile					0.419** [0.310, 0.568]
Constant	4.692** [3.590, 6.133]	4.968** [3.737, 6.603]	4.096** [1.799, 9.326]	4.492** [1.900, 10.624]	9.218** [3.718, 22.852]
Observations	2,137	2,137	2,137	2,137	2,137
Log likelihood	-1,120.558	-1,119.845	-1,119.276	-1,116.045	-1,098.141
AIC	2,251.116	2,251.691	2,256.551	2,256.090	2,222.282

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A19 Stepwise regression results for Germany, outcome: provided e-mail address

Stepwise regression results for Germany, outcome: provided e-mail address					
	Odds ratio estimate for binary logistic regression, Germany				
	Outcome: provided e-mail address (ref. did not provide e-mail address)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	0.616** [0.483, 0.786]	0.547** [0.427, 0.703]	0.564** [0.438, 0.726]	0.552** [0.426, 0.714]	0.580** [0.447, 0.752]
Design: climate-loose	0.842 [0.638, 1.112]	0.823 [0.622, 1.088]	0.837 [0.633, 1.109]	0.810 [0.608, 1.078]	0.856 [0.642, 1.143]
Design: immigration-strong	0.504** [0.395, 0.643]	0.464** [0.363, 0.595]	0.471** [0.368, 0.604]	0.481** [0.374, 0.620]	0.503** [0.390, 0.649]
Design: immigration-loose	0.894 [0.682, 1.171]	0.815 [0.620, 1.072]	0.819 [0.622, 1.077]	0.803 [0.607, 1.062]	0.830 [0.627, 1.100]
Sex: male		1.539** [1.300, 1.821]	1.532** [1.294, 1.814]	1.436** [1.207, 1.708]	1.385** [1.163, 1.649]
Age: 40–59			0.868 [0.586, 1.285]	0.908 [0.611, 1.350]	0.905 [0.608, 1.348]
Age: 60–64			0.910 [0.602, 1.377]	0.983 [0.646, 1.495]	0.983 [0.645, 1.497]
Age: 65+			1.015 [0.681, 1.512]	1.135 [0.757, 1.701]	1.138 [0.758, 1.710]
Income: low				0.920 [0.734, 1.152]	0.891 [0.710, 1.117]
Income: medium				0.920 [0.744, 1.138]	0.916 [0.740, 1.134]
Income: missing				0.313** [0.241, 0.407]	0.304** [0.233, 0.396]
Device: mobile					0.556** [0.414, 0.746]
Constant	1.090 [0.909, 1.308]	0.907 [0.745, 1.104]	0.964 [0.633, 1.470]	1.172 [0.755, 1.818]	1.994** [1.190, 3.341]
Observations	2,455	2,455	2,455	2,455	2,455
Log likelihood	-1,661.423	-1,648.694	-1,647.303	-1,601.673	-1,593.934
AIC	3,332.845	3,309.389	3,312.606	3,227.346	3,213.869

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

Table A20 Stepwise regression results for the United States, outcome: provided e-mail address

Stepwise regression results for the United States, outcome: provided e-mail address					
	Odds ratio estimate for binary logistic regression, United States				
	Outcome: provided e-mail address (ref. did not provide e-mail address)				
	(1)	(2)	(3)	(4)	(5)
Design: climate-strong	0.854 [0.669, 1.090]	0.790† [0.614, 1.018]	0.841 [0.651, 1.086]	0.895 [0.687, 1.166]	0.983 [0.751, 1.285]
Design: climate-loose	0.722* [0.555, 0.939]	0.716* [0.550, 0.932]	0.776† [0.594, 1.013]	0.788† [0.598, 1.038]	0.860 [0.651, 1.138]
Design: immigration-strong	0.661** [0.513, 0.853]	0.648** [0.502, 0.836]	0.689** [0.533, 0.891]	0.725* [0.556, 0.946]	0.737* [0.564, 0.964]
Design: immigration-loose	0.847 [0.637, 1.125]	0.807 [0.605, 1.076]	0.864 [0.647, 1.156]	0.890 [0.659, 1.200]	1.025 [0.755, 1.392]
Sex: male		1.221* [1.036, 1.440]	1.200* [1.017, 1.416]	1.067 [0.897, 1.271]	0.996 [0.834, 1.188]
Age: 40–59			0.399† [0.155, 1.026]	0.370* [0.141, 0.972]	0.357* [0.136, 0.940]
Age: 60–64			0.290** [0.114, 0.737]	0.273** [0.105, 0.710]	0.260** [0.100, 0.677]
Age: 65+			0.258** [0.106, 0.627]	0.276** [0.111, 0.683]	0.271** [0.109, 0.672]
Income: low				1.422* [1.074, 1.884]	1.429* [1.077, 1.895]
Income: medium				1.023 [0.807, 1.297]	1.033 [0.814, 1.311]
Income: missing				0.370** [0.287, 0.478]	0.378** [0.292, 0.488]
Device: mobile					0.583** [0.469, 0.724]
Constant	1.491** [1.227, 1.812]	1.379** [1.123, 1.693]	4.864** [1.997, 11.846]	5.880** [2.310, 14.967]	8.955** [3.456, 23.204]
Observations	2,571	2,571	2,571	2,571	2,571
Log likelihood	-1,765.883	-1,763.050	-1,754.446	-1,678.126	-1,665.927
AIC	3,541.767	3,538.100	3,526.891	3,380.252	3,357.853

Notes: AIC = Akaike information criterion. †$p < .10$, *$p < .05$, **$p < .01$. Unweighted. Values in brackets refer to the 95% confidence interval.

A4 R Packages Used

Package	Version	Citation
arm	1.13.1	Gelman and Su (2022)
base	4.3.2	R Core Team (2023)
ggpubr	0.6.0	Kassambara (2023)
here	1.0.1	Müller (2020)
Hmisc	5.1.1	Harrell Jr (2023)
janitor	2.2.0	Firke (2023)
kableExtra	1.4.0	Zhu (2024)
psych	2.4.1	Revelle (2024)
stargazer	5.2.3	Hlavac (2022)
tidyselect	1.2.0	Henry and Wickham (2022)
tidyverse	2.0.0	Wickham et al. (2019)

Package Citations

Firke, S. (2023). *janitor: Simple tools for examining and cleaning dirty data*. https://CRAN.R-project.org/package=janitor

Gelman, A., & Su, Y. (2022). *arm: Data analysis using regression and multilevel/hierarchical models*. https://CRAN.R-project.org/package=arm

Harrell, F. E. (2023). *Hmisc: Harrell miscellaneous*. https://CRAN.R-project.org/package=Hmisc

Henry, L., & Wickham. H. (2022). *tidyselect: Select from a set of strings*. https://CRAN.R-project.org/package=tidyselect

Hlavac, M. (2022). *stargazer: Well-formatted regression and summary statistics tables*. Social Policy Institute. https://CRAN.R-project.org/package=stargazer

Kassambara, A. (2023). *ggpubr: "ggplot2" based publication ready plots*. https://CRAN.R-project.org/package=ggpubr

Müller, K. (2020). *here: A simpler way to find your files*. https://CRAN.R-project.org/package=here

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rodriguez-Sanchez F, & Jackson, C. (2024). *_grateful: Facilitate citation of R packages.* https://pakillo.github.io/grateful

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V. (...), Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org

Zhu, H. (2024). *kableExtra: Construct complex table with "kable" and pipe syntax*. https://CRAN.R-project.org/package=kableExtra

methods, data, analyses | Vol. 19(1), 2025, pp. 137-164 DOI: 10.12758/mda.2025.09

Improving Understanding of Survey Questions with Multimodal Clarification

Maura Spiegelman[1] & Frederick Conrad[2]

[1] *Independent Researcher*
[2] *University of Michigan*

Abstract

If survey respondents do not interpret a question as it was intended, they may, in effect, answer the wrong question, increasing the chances of inaccurate data. Researchers can bring respondents' interpretations into alignment with what is intended by defining the terms that respondents are at risk of misunderstanding. This article explores strategies to increase alignment between researchers' intentions and respondents' answers by taking advantage of the unique affordances of online surveys compared to paper or other analog formats. Web surveys are often text-based, but allow for the seamless integration of embedded audio material so that users may read, listen to, or both read and listen to survey instructions. Unimodal definitions are either spoken or textual, while multimodal definitions are both spoken and textual. Further, definitions can be designed to take advantage of the affordances of each mode. While *mode-invariant* definitions contain the same words irrespective of whether they are textual or spoken, *mode-optimized* definitions are designed to take advantage of the affordances of written and spoken communication. For example, definitions optimized for textual presentation use fewer words than corresponding mode-invariant definitions and are designed so the key information is visually salient, while definitions optimized for spoken presentation are shorter and more colloquial than corresponding mode-invariant definitions. In this study, both mode-optimized and mode-invariant formats improved alignment. Multimodal, mode-optimized definitions produced improved alignment over both types of unimodal definitions. This study suggests that multimodal definitions, when thoughtfully designed, can improve data quality in online surveys without negatively impacting respondents.

Keywords: question definitions, questionnaire instructions, audio input, multimodal input, data quality, web survey

Ensuring that survey respondents interpret survey questions as their authors intended is a prerequisite for producing high quality data. Otherwise, respondents may, in effect, answer a different question than the one the researchers believed they were asking, potentially resulting in inaccurate answers. One way to align respondents' and researchers' interpretations is to clarify terms that may not map cleanly to respondents' circumstances. For example, if a respondent is unsure whether to include TV programming streamed to their laptop computer when answering a question about their recent TV watching, defining exactly what is meant by TV watching should resolve the respondent's uncertainty about how to answer. Explicitly clarifying terms can help assure that respondents understand survey questions—whether asked by interviewers or self-administered—as intended and in a way that fits their situation. In everyday conversation, participants *ground* what has been said (Clark, 1996) by discussing the speaker's intentions until both parties agree they understand each other well enough to accomplish the goals of the conversation. The benefits of grounding meaning have been explored in survey interviews, self-administered online questionnaires, virtual interviews, and speech dialog systems (see Conrad & Schober (2021) for a summary and review). This prior research concerns the delivery of unimodal, that is, solely spoken or solely textual definitions. However, there may be value in exploring multimodal delivery of definitions. In educational psychology, researchers have found that multimodal communication can improve comprehension and information retention compared to unimodal communication in some, though not all, circumstances (Moreno & Mayer, 2002; Mousavi et al., 1995). This paper builds upon the research in both conversational grounding and multimodal communication to explore whether multimodal definitions, that is, definitions that are both spoken and textual, can improve the quality of survey responses relative to unimodal definitions (either spoken or textual) or no clarification. This study also tests the conditions under which multimodal definitions might be most effective, that is, multimodal definitions that are identically worded across the two modes or that exploit the affordances of each mode in which they are implemented.

Notes

This work was conducted at the Joint Program in Survey Methodology, University of Maryland, USA and was supported a Rensis Likert Dissertation Research Award, University of Michigan. The authors declare there is no conflict of interest.

Direct correspondence to

Maura Spiegelman
E-mail: mspiegelman@gmail.com

Survey Definitions

Surveys ask respondents about conditions or situations of varying complexity, clarity, and familiarity. When respondents' understanding of ideas or terms is different from what researchers intend, data quality is likely to suffer unless understanding can be aligned (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober et al., 2018). For example, the concept of how many people live in a household is straightforward for respondents in most living situations. However, for respondents with a child living away at college, it is not clear whether they should include their child in their response, potentially introducing bias if such misalignments occur in one direction. For the portion of respondents who have children living at college, a definition of who should be counted as living in the household can correct respondents' misconceptions, leading to an accurate answer.

Although it can improve comprehension of questions and response accuracy, providing definitions generally increases the amount of time needed to answer survey items (Conrad & Schober, 2020, 2001; Conrad et al., 2007; Schober et al., 2004; Schober & Conrad, 1997), although West et al. (2018) found no effect on response times. Respondents need to listen to or read the definitions and incorporate them into their interpretation of the question—which could potentially reduce their satisfaction with the interview experience, potentially reducing completion rates, and likely inflating sampling variance.

While definitions can certainly help align respondents' and researchers' understanding of survey questions, providing definitions will only provide these benefits if respondents use them. One reason respondents might not use a definition is if it is hard to understand. This might be the case if, for example, the definition is presented in textual form and the respondent is not a strong reader, or because a spoken definition is so complicated and long that a momentary lapse in the respondent's attention might result in their not understanding the definition. The first of these might be addressed by presenting the definition in both textual and spoken forms, a multimodal definition. Not only might this increase the chances that the content of the definition is interpretable by most respondents, but it emphasizes the importance of the definition by conveying it in two ways. The second issue, that the definition is long and complicated, might be addressed through improved design, such as simplifying the content and presenting the definition in a way that is most appropriate to its mode.

Multimodal Communication

Multimodal communication typically involves the simultaneous presentation of information in two or more channels of communication. In the case of survey questionnaires this can involve the way content such as questions and definitions is presented, how respondents report their answers, or both (Johnston,

2008). For example, online data collection can be designed so that respondents can both read a textually presented question and hear the corresponding spoken question; or to enable respondents to answer by either typing/clicking or speaking (e.g., "Type 1 or say 'Yes'"). This study explores the former: multimodal presentation of information, in particular, definitions of survey concepts.

Research on multimodal communication has combined spoken information with a variety of visual presentations and has found that, under the right circumstances, combining audio and visual material can be more effective than using only one mode. For example, in educational psychology, researchers observed that presenting students with both audio and visual material is more effective pedagogically than using only spoken communication for certain types of information and presentations. For example, Moreno and Mayer (2002) noted that participants showed higher levels of retention and were more effective at applying information in a new context (rather than simply recalling it) when taught in a multimodal, rather than unimodal spoken format. Mousavi, Low, and Sweller (1995) found that students needed less time to accurately solve geometry problems using combined diagrams—which are visual—with orally presented verbal information rather than textually presented verbal material that competes with diagram processing for limited visual attention. By comparing sequential and simultaneous presentation, they attributed these results to the relatively low cognitive load of using multiple communication channels, due to partial independence of visual and verbal processing (Mayer, 2014; Sweller et al., 2019).

Extending these findings to processing survey questions, multimodal presentation could help respondents understand and apply survey definitions, presumably improving the quality of their answers. By dividing content between the textual and spoken material, the amount of content presented in either mode is reduced relative to unimodal presentation. This division by mode is particularly helpful for processing spoken information which is ephemeral and (unless it is audio-recorded and can be replayed) will be gone after it is presented. In contrast, textual information is persistent, (i.e., remains visible over time) and can be read while the spoken information is presented, after it is presented, or both. Thus, textual information does not need to be stored in working memory initially, the way spoken information does, because it is preserved externally (i.e., on the screen). Moreover, working memory is hypothesized to consist of separate storage mechanisms for textual ("visuo-spatial") and spoken ("phonological") information, coordinated by a central executive system (Baddeley, 1992; also see Dumas et al., 2009), suggesting that it is well suited to multimodal presentation of information.

However, in the psychological literature, the benefits of multimodal communication depend on the extent to which the information in the different modes is redundant and conveys the same information. When information is simultaneously conveyed both orally and textually, redundancy can potentially reduce

comprehension. For example, when an animated technical explanation was combined with either spoken narration only or identical, simultaneous spoken and textual narration, the latter treatment resulted in poor comprehension, evident in reduced retention and transfer of information (Mayer et al., 2001). That is, participants who were exposed to redundant spoken and written words and a complementary animation had lower comprehension than participants who were exposed to only spoken words and a complementary animation. Note that many of these studies involve visual stimuli that created substantial cognitive demands, such as combinations of written instructional text, numerical tables, and graphs or diagrams (e.g., Kalyuga et al., 2004). Since survey researchers generally attempt to convey less complicated material to respondents, rarely requiring animated instruction, these findings are unlikely to limit the effectiveness of multimodal material in surveys, but they do point out that redundant content across modes can degrade respondents' ability to process additional information.

When spoken and textual content does not consist of exactly the same words but instead conveys the same underlying message, this kind of semantic redundancy does not seem to harm comprehension the way literal redundancy does (Kalyuga et al., 2004). Mild levels of redundancy, for example key words or phrases, have been shown to increase retention (Mayer & Johnson, 2008). This suggests that multimodal definitions of survey concepts can yield higher rates of comprehension when the text emphasizes somewhat different ideas than the spoken content, rather than simply duplicating the information. More specifically, identical spoken and textual definitions may reduce comprehension, while complementary definitions seem likely to improve comprehension.

This study tests whether multimodal definitions for key concepts in survey questions can improve the quality of responses (i.e., their alignment with definitions) compared to unimodal (either spoken or textual) definitions. Two types of multimodal definitions were tested: mode-invariant definitions, with fully redundant spoken and textual information (i.e., the same *words* presented visually and via speech) and mode-optimized definitions, designed specifically for each mode with partially redundant content, i.e., the same *concepts* conveyed textually and orally using complementary wording and exploiting the affordances of each mode. If both types of multimodal definitions outperform unimodal definitions, this would likely be due to the partial independence of communication channels. If only mode-optimized multimodal definitions outperform unimodal definitions, this would likely be due to the literal redundancy of mode-invariant definitions, which provide no additional information or perspective on the underlying concept from their multimodality, even potentially interfering with comprehension.

Methods

Experimental Design

Respondents completed a web survey in one of seven experimental conditions distinguished by the type of definitions made available: (1) none (i.e., the control condition), (2) textual, mode-invariant, (3) textual, mode-optimized, (4) spoken, mode-invariant, (5) spoken, mode-optimized, (6) multimodal (i.e., both textual and spoken), mode-invariant), or (7) multimodal (i.e., both textual and spoken), mode-optimized. Irrespective of the mode(s) and optimization of definitions, all survey questions were presented textually. These seven conditions comprise a fraction of all possible combinations of mode, multimodality, and format, but allow for the most important comparisons: whether multimodal definitions were more effective, i.e., promoted greater alignment of respondents' understanding of each question with the question's intended interpretation, than unimodal definitions and whether mode-optimized, multimodal definitions—in which the information in each mode was complementary and relatively non-redundant—were more effective than mode-invariant, multimodal definitions.

Respondents were asked to provide numeric responses to 15 survey questions, each accompanied by a definition for the key concept (except in the control condition). Definitions were either "inclusive" (five questions) or "exclusive" (seven questions). Inclusive definitions were designed to expand the scope of what behaviors could be counted as examples of the concept in question (for example, *including* commuting when reporting the amount of work for which the respondent was paid), and exclusive definitions were designed to reduce the scope (for example, *excluding* streamed or recorded content when reporting on the amount of television watched; Schober & Conrad, 2000). To promote the questionnaire's coherence, questions on similar topic areas were grouped together, for example, hours spent watching television and listening to the radio. Thus, all respondents viewed questions in the same order. Finally, all respondents were asked a series of debriefing questions about their demographics and experience during the study. All surveys were identical except for the type of definition made available.

Mode-invariant definitions were designed to emulate the format of data collection instruments from many government statistical agencies. The definitions in these types of surveys contain detailed information, and when presented in textual format, often appear as a dense paragraph; they are not designed for respondents to identify the subcomponents most relevant to their situations. When these same definitions are read aloud, they do not flow like a conversation or other spoken communication. Instead, the experience is reminiscent of questionnaires designed to be self-administered (either on paper or online) but which are administered by an interviewer over the phone to some respondents. For multimodal mode-invariant definitions, identical wording was used for both the spoken and textual components leading to fully redundant information. For

multimodal mode-optimized definitions, spoken optimized and textual optimized components were presented together.

Mode-optimized definitions were designed to be easier for respondents to either read or listen to and to help respondents identify relevant information by following best practices of written and spoken communication. For textual mode-optimized definitions, factors known to facilitate text comprehension (White, 2012) were used: bolded text to draw attention to key words and phrases, bullets and other organizational devices to divide text into logical groupings. For each question, mode-optimized textual definitions had lower Flesch-Kincaid grade level reading scores (shorter sentence length and fewer syllables per word) than their mode-invariant counterparts (Flesch, 1948).

For spoken mode-optimized definitions, the scripts were designed to follow best practices for spoken communication. For example, in order to facilitate comprehension in spoken mode-optimized definitions, extraneous information that was included in their mode-invariant counterparts was removed (Sweller et al., 1990). Shorter spoken definitions are also less taxing on respondents' working memory, and require relatively little effort to comprehend compared to longer, mode-invariant definitions (Leahy & Sweller, 2011). For each question, mode-optimized spoken definitions were shorter in duration than their mode-invariant counterparts (an average of 11.4 seconds compared to 23.1 seconds). Spoken mode-optimized definitions were read aloud, audio-recorded, and played back by the researchers to judge their flow and ease of comprehension, then adjusted iteratively, if needed in the researchers' judgment. The displayed text and scripts for all mode-invariant and mode-optimized definitions are shown in the Appendix and screenshots of each condition are shown in Table 1.

Data Collection

We implemented the experimental conditions—the seven different web-based questionnaires—in Qualtrics, using TurkPrime (now CloudResearch) to recruit and manage participants from Amazon Mechanical Turk (MTurk). Each condition was posted as a separate "task" within MTurk, with identical descriptions, and participants were only eligible to complete one of these tasks, essentially randomizing respondents across treatment groups. A $1 incentive was provided to respondents upon completion of the survey. The median completion time for the surveys, including debriefing and other non-experimental questions, was just under 10 minutes, resulting in a median hourly rate of $6.17. The University of Maryland Institutional Review Board approved this study, and we collected data in the summer of 2018.

In total, 1,014 respondents completed the study. For the 12 experimental survey questions, 11,988 total observations were retained for analysis after removing impossible and implausible values (for example, reports of participating in

any given activity for close to 168 hours per week since there are, in total, only 168 hours per week). The distribution of respondents and observations by experimental condition is shown in Table 2.

Table 1 Selected screenshots by experimental condition

Experimental condition	Screenshot
Control	In the past 7 days, how many hours of television did you watch?
Spoken mode-invariant	In the past 7 days, how many hours of television did you watch? Play for more information: 
Spoken mode-optimized	In the past 7 days, how many hours of television did you watch? Play for more information: 
Textual mode-invariant	In the past 7 days, how many hours of television did you watch? Watching television includes any programs other than films. This may include sitcoms, dramas, news, sports, and reality shows. Television is watched on a television set at the time it is broadcast and does not include programming recorded with a DVR, viewed on-demand, or streamed. Include content viewed on a television set only and exclude any content viewed on a computer or mobile device.
Textual mode-optimized	In the past 7 days, how many hours of television did you watch? • **Content is broadcast.** Exclude DVRed, on-demand, and streamed shows. • **TV set.** Exclude shows watched on a computer or mobile device. • **TV shows.** Exclude films, even if watched while they air.

Table 1 (continued)

Multimodal mode-invariant	In the past 7 days, how many hours of television did you watch? Play for more information:  Watching television includes any programs other than films. This may include sitcoms, dramas, news, sports, and reality shows. Television is watched on a television set at the time it is broadcast and does not include programming recorded with a DVR, viewed on-demand, or streamed. Include content viewed on a television set only and exclude any content viewed on a computer or mobile device. ☐
Multimodal mode-optimized	In the past 7 days, how many hours of television did you watch? Play for more information:  • **Content is broadcast.** Exclude DVRed, on-demand, and streamed shows. • **TV set.** Exclude shows watched on a computer or mobile device. • **TV shows.** Exclude films, even if watched while they air. ☐

Table 2 Sample sizes and number of observations by experimental condition

	# Respondents	# Observations
Control	104	1,239
Spoken mode-invariant	200	2,356
Spoken mode-optimized	162	1,920
Textual mode-invariant	101	1,196
Textual mode-optimized	80	952
Multimodal mode-invariant	160	1,890
Multimodal mode-optimized	207	2,435
Total	1,014	11,988

The number of respondents and observations varies by condition for two primary reasons. First, a subset of participants were recruited for a pilot test with control, spoken mode-invariant, textual mode-invariant, and multimodal

mode-optimized definitions. When it was evident that the procedure worked as expected, these cases were pooled with newer cases. In addition, more respondents were recruited into treatment groups with spoken components (both unimodal and multimodal) under the assumption that not all respondents would comply with instructions and play the spoken definitions.

Analytic Strategy

Alignment of Question Interpretation and Intended Meaning

Because different questions asked about different reference periods and different types of activities and measured the target behavior on different scales, responses could not be averaged in raw form. Moreover, for some questions—namely those for which definitions were inclusive—higher numeric responses indicated consistency with definitions, while for other questions—namely those for which definitions were exclusive—lower numeric responses indicated consistency with definitions. So that we could compare across questions and conditions, we converted responses to each question to a z-score, trimmed to +4 and -4, and then multiplied these z-scores by -1 for questions with exclusive definitions. Because of this trimming, the mean z-score per question deviates slightly from 0. This conversion allows responses to be pooled across questions, using a standard scale, and for results from each condition to be pooled, namely higher values indicate greater alignment with definitions (more standard deviations from the mean response) while lower values indicate that responses are less aligned with definitions, irrespective of whether a definition was inclusive or exclusive.

We used a general linear mixed model in SAS 9.4 to compare the effects of different definition treatments by modeling z-scores while accounting for clustering of observations within respondents. Questions (denoted with subscript q) are nested within respondents (denoted with subscript i). Both questions and respondents were given random intercepts, allowing for baseline differences in question difficulty and respondents' behavior, though respondents are treated as random effects and questions as fixed effects.

$$Y_{iq} = \gamma_{00} + \gamma_{10}\,(D_i) + U_{i0} + e_{iq}$$

We conducted F-tests at the observation level, rather than the respondent level, when analyzing alignment of responses with definitions (D_i) using type 3 F-tests of fixed effects unless otherwise stated. For pairwise comparisons of point estimates, we use Student's t-tests unless otherwise stated.

Respondent Use of Definitions

Some respondents did not fully comply with the instructions to attend to definitions, raising the possibility that the effect of definition type on alignment might be stronger for those who comply. To address this, observations can be compared both overall and by examining only observations based on (inferred) compliance with the experimental treatment. For spoken definitions, we captured whether audio clips were fully played, and for textual definitions, we estimated the time that would be required to read a particular question and its associated definition and compared this to the actual time each respondent spent on the page. Note that compliance is relevant for the control group even though control respondents were not provided with any definitions; these respondents were expected to spend sufficient time on each page to read survey questions.

To measure respondents' exposure to spoken definitions, the online survey captured how many times a spoken definition was fully played by using embedded JavaScript code. It is important to note that this measure could not record whether a respondent's audio was muted, nor whether they truly attended to the spoken information, but instead serves as a proxy for respondent compliance in a self-interview setting that involved auditory information. Respondents could play a definition by clicking the "play" icon (right arrow) in a standard media bar. For spoken definitions, whether as part of a unimodal or multimodal format, a given response was considered "compliant" if the audio file was fully played.

We inferred whether respondents read the textual information they were presented by determining if the time spent on any given question was at least as long as the estimated reading time for that text, in which case they were determined to have complied with instructions. We calculated the reading time threshold for each question and each textual definition (in the relevant conditions) by counting the words and multiplied the counts by 200 msec. This is the average reading speed, according to Carver (1992), for adult Americans when reading to retain content for relatively short intervals, as is needed when answering survey questions[1]. Thus, the word count for the control group and groups with unimodal spoken definitions were identical (question word count only), and the word count for the textual only and multimodal definition conditions were identical for mode-invariant versions (question and definition word count) as was also the case for the mode-optimized versions (question and definition word count). However, like the proxy for spoken definition compliance, this criterion does

[1] Conrad et al. (2017) and Zhang and Conrad (2014) used 300 msec/word for similar purposes. However, their thresholds were intended to account for reading plus thinking time, so faster responses could be considered speeding. Because the tasks in our study were less cognitively burdensome, we selected a more conservative threshold in order to avoid inflating our estimates of the impact of compliance on alignment of question interpretation and intended meaning.

not guarantee that respondents truly attended to and absorbed the textual information presented to them. Eye-tracking could help determine whether respondents viewed the textual information, for example, whether they fixated on textual information in left-to-right, top-to-bottom order or whether they skipped or sped through information. However, even knowing what they looked at would not capture whether they deeply comprehended and internalized the information or merely scanned the text. In a self-interview setting, time per page is the best available measure of reading time and thus proxy for respondent compliance. Note that for both spoken and textual definitions, compliance was treated as a binary metric for which a given observation either met compliance criteria or did not.

Results

Alignment of Question Interpretation and Intended Meaning

The average *z*-scores by mode and optimization of definitions are shown in Table 3. *Z*-scores indicate the number of standard deviations by which observations for a given definition type varied from the average response across all questions and definition modes. Higher values indicate more alignment with definitions, while lower values indicate less alignment with definitions. For example, the average *z*-score for responses to questions in the control group was about -0.13, indicating that those responses were less aligned with definitions than the average response by 0.13 standard deviations.

Table 3 Mean *z*-score by definition mode and optimization

Definition mode	Optimization	Mean *z*-score
Control (no definition)	n/a	-0.126
Spoken	All	-0.009
	Mode-invariant	-0.025
	Mode-optimized	0.012
Textual	All	-0.014
	Mode-invariant	-0.032
	Mode-optimized	0.008
Multimodal	All	0.041
	Mode-invariant	0.013
	Mode-optimized	0.063

Responses to questions with multimodal definitions were more aligned with definitions than the average response by about 0.041 standard deviations, and they were significantly more aligned than responses to questions with only unimodal definitions. That is, average *z*-scores were higher for the multimodal group than unimodal textual definitions ($t(996) = 2.33$, $p = .020$), and unimodal spoken definitions ($t(1002) = 2.56$, $p = .011$). Overall, definition mode was a significant predictor of the degree to which responses were aligned with definitions ($F(3,994) = 11.37$, $p < .001$). As expected, responses were least aligned with definitions for the control group, under which no definitions were available.

The effectiveness of multimodal definitions appears to be driven by their optimization. That is, average *z*-scores were higher for the multimodal mode-optimized definitions than for unimodal mode-invariant definitions, both spoken ($t(1003) = 3.39$, $p < .001$) and textual $t(997) = 2.98$, $p = .003$). *Z*-scores for multimodal mode-invariant definitions were higher, though not significantly so, than *z*-scores for unimodal mode-optimized definitions both spoken ($t(997) = 1.87$, $p = .062$) and textual ($t(990) = 1.59$, $p = .111$). They were marginally higher than for multimodal, mode-invariant definitions ($t(1000) = 1.80$, $p = .072$) making it somewhat ambiguous to what extent the presence alone of complimentary, rather than redundant, multimodal information can improve data quality.

Respondents' Use of Definitions

Compliance Rates

For the four definition types with a spoken component (spoken mode-invariant, spoken mode-optimized, multimodal mode-invariant, multimodal mode-optimized), compliance with spoken definitions (that is, fully playing a definition's audio file) ranged from the relatively low rate of 29% for multimodal mode-invariant definitions to 47% for spoken mode-optimized definitions (Table 4).

For all four treatment groups in which spoken definitions were available, compliance was highest with the first definition presented in the survey, ranging from 61% for multimodal mode-invariant to 84% for multimodal mode-optimized. Across all spoken mode conditions, the compliance rate for the first survey question was significantly higher than the compliance rate for every other question at the $p < .05$ level. A steady decline in compliance might reflect respondent fatigue; the reason for this abrupt drop is unclear but could have occurred if respondents noted there were no direct repercussions of answering without playing the entire spoken definition. This drop-off in compliance occurred both overall and within each of the 4 treatment groups with spoken definitions. For both mode-invariant and mode-optimized definitions, compliance with instructions to play the audio was higher for respondents who only received spoken definitions, rather than multimodal respondents who were encouraged to both read and listen to definitions. Compliance was higher for the spoken mode-optimized

group than for either of the multimodal conditions, and higher for the spoken mode-invariant than multimodal mode-invariant group. However, it should be noted that these overall compliance rates were less than 50% for each condition.

Differences in compliance between unimodal and multimodal groups may be driven by the presence of an alternative way of acquiring multimodal definition content. For respondents in unimodal spoken groups who were inclined to use definitions in their responses, their only choice was to listen to spoken definitions. Respondents in multimodal groups could have given responses consistent with definitions by reading textual definitions, even if they did not fully play an audio clip.

Table 4 Compliance with spoken and textual portions of definitions by question number and definition type

Question		Textual mode-invariant	Textual mode-optimized	Spoken mode-invariant	Spoken mode-optimized	Multimodal mode-invariant	Multimodal mode-optimized
1	Listened	n/a	n/a	73%	78%	61%	84%
	Read	55%	88%	n/a	n/a	97%	99%
2	Listened	n/a	n/a	59%	54%	40%	42%
	Read	21%	40%	n/a	n/a	58%	58%
3	Listened	n/a	n/a	41%	50%	32%	32%
	Read	23%	76%	n/a	n/a	53%	84%
4	Listened	n/a	n/a	44%	52%	30%	38%
	Read	42%	61%	n/a	n/a	54%	74%
5	Listened	n/a	n/a	39%	46%	27%	38%
	Read	31%	66%	n/a	n/a	50%	82%
6	Listened	n/a	n/a	36%	45%	26%	26%
	Read	36%	70%	n/a	n/a	62%	82%
7	Listened	n/a	n/a	26%	33%	24%	26%
	Read	21%	64%	n/a	n/a	44%	84%
8	Listened	n/a	n/a	34%	40%	22%	29%
	Read	29%	64%	n/a	n/a	49%	74%
9	Listened	n/a	n/a	37%	38%	26%	25%
	Read	33%	74%	n/a	n/a	50%	74%
10	Listened	n/a	n/a	40%	46%	20%	35%
	Read	29%	59%	n/a	n/a	42%	67%
11	Listened	n/a	n/a	25%	39%	21%	26%
	Read	18%	48%	n/a	n/a	41%	67%
12	Listened	n/a	n/a	31%	40%	21%	25%
	Read	17%	70%	n/a	n/a	43%	78%
Overall	Listened	n/a	n/a	39%	47%	29%	35%
	Read	29%	65%	n/a	n/a	53%	78%

Comparing overall compliance for mode-optimized and mode-invariant definitions, the rate was higher for respondents in the former than latter (47% and 39%, respectively, for unimodal; 35% and 29%, respectively, for multimodal), though this difference was only significant when comparing the two unimodal conditions.

Compliance with textual definitions followed a similar pattern. Again, this type of compliance was operationalized as at least as much time spent on a given question as the estimated reading time for the question and definition text. Compliance was significantly higher for the first question than every other subsequent question ($p < .001$) for each definition type, similar to the pattern shown for compliance with spoken definitions (see Table 4). In addition, compliance rates differed by condition for each pairwise comparison between the 4 groups with textual definitions. In particular, the 78% compliance rate for multimodal mode-optimized definitions was significantly higher than the 65% compliance rate for textual mode-optimized definitions ($t(544) = 3.54$, $p < .001$), which was significantly higher in turn than the 53% compliance rate for multimodal mode-invariant definitions ($t(544) = 2.74$, $p = .006$), which was significantly higher than the 29% compliance rate for textual mode-invariant definitions ($t(544) = 6.54$, $p < .001$). So, compliance was highest for mode-optimized definitions. For both mode-invariant and mode-optimized definitions, compliance was higher for multimodal than unimodal definitions. However, as with spoken definitions, all mode-optimized textual definitions had fewer words than all mode-invariant textual definitions, and presumably as a result, shorter estimated reading times. As a result, length and optimization are confounded and prevent us from distinguishing the effects of optimization per se from reduced text on compliance.

Compliance with multimodal definitions depends on whether respondents only read, only listened to, or both read and listened to definitions. However, in this study it is important to note that the duration of each spoken definition was at least as long as the estimated reading time for the corresponding textual definition, so all respondents who fully listened to a multimodal definition's spoken component were coded as being in full compliance.

Alignment of Question Interpretation and Intended Meaning When Respondents Access Definitions

We have suggested that definitions—in either mode—can help align respondents' understanding of questions with the questions' intended meaning, However, increased alignment could be due to the mere presence of the definitions rather than the content of the definitions. For example, multimodal definitions may signify to respondents that the information is important, or because the content of the definitions is better understood by respondents. These possible explanations cannot be disentangled without examining responses while considering whether individuals accessed the definitions that were available to them.

We can treat noncompliant responses in two different ways. They may be dropped from analysis, for example, a response to a textual definition that did not meet the criteria for reading can be omitted entirely. Alternatively, that response effectively had the same de facto treatment as a control group response and could be analyzed with the others from that group, though, respondents may have read part of a definition or read all text more quickly than the estimated reading speed threshold, so their experiences may not be identical to those of actual control group participants. Any observations for which a respondent in a multimodal condition did not both fully listen to and read the definition can be analyzed with the control, unimodal textual or unimodal spoken groups (although the latter is theoretical given that audio clips were longer than estimated reading duration so playing an audio file will lead to a compliant reading classification even if the respondent did not read the definition). As with spoken definitions, observations were categorized based on whether they fully met compliance criteria, so observations for which definitions may have been partially played or read were considered noncompliant and analyzed accordingly. Observations that did not meet criteria for the control group, that is, the amount of time spent on the page was less than the compliance cutoff for fully reading the question text, were excluded since they could not be treated as compliant with any treatment group. Table 5 shows the average *z*-score for both methods of categorizing compliant and noncompliant responses.

Table 5 Mean *z*-score by definition mode and optimization for compliant responses and de facto treatment

Definition mode	Optimization	Mean *z*-score (compliant responses only)	Mean *z*-score (by de facto treatment)
Control (no definition)	n/a	-0.129	-0.071
Spoken	All	0.076	0.075
	Mode-invariant	0.073	0.071
	Mode-optimized	0.079	0.079
Textual	All	0.059	0.076
	Mode-invariant	0.018	0.041
	Mode-optimized	0.082	0.093
Multimodal	All	0.163	0.163
	Mode-invariant	0.079	0.079
	Mode-optimized	0.217	0.217

When limiting analysis to only observations that met compliance criteria, responses to questions with multimodal definitions were more aligned with definitions than the average response by about 0.16 standard deviations (Table 5).

That is, average z-scores were higher for definitions that were multimodal than unimodal textual ($t(691) = 2.86$, $p = .004$) and unimodal spoken ($t(619) = 2.78$, $p = .006$) definitions.

As with the analysis of all observations (irrespective of compliance), this difference is driven by multimodal mode-optimized definitions. Responses to these questions were more aligned with the underlying concepts than the average response by about 0.22 standard deviations. That is, average z-scores were higher for the mode-optimized multimodal group than for all other conditions: spoken mode-invariant ($t(649) = 3.41$, $p < .001$), spoken mode-optimized ($t(601) = 3.26$, $p = .001$), textual mode-invariant ($t(959) = 3.46$, $p < .001$), textual mode-optimized ($t(614) = 2.97$, $p < .001$), and multimodal mode-invariant groups ($t(643) = 2.78$, $p = .006$) groups. The increased data quality with multimodal definitions is primarily attributed to presenting complementary, rather than redundant, information.

Looking at de facto treatment, that is, categorizing responses based on the treatment they effectively received rather than the group to which they were originally assigned, we observed a similar pattern. Answers reported when respondents were compliant with multimodal definitions were significantly more aligned with definitions than each of the other de facto definition types. That is, average z-scores were higher for the multimodal group than when the effective treatment was textual definitions ($t(2567) = 2.89$, $p = .004$), spoken definitions ($t(1517) = 2.88$, $p = .004$), and the control treatment with no definitions ($t(1713) = 8.99$, $p < .001$). Overall, de facto definition mode significantly predicted the degree to which responses were aligned with definitions ($F(3,2057) = 37.57$, $p < .001$).

Observations produced when respondents complied with multimodal mode-optimized definitions were significantly more aligned with definitions than each of the other types of definition. That is, with de facto categorization, average z-scores were higher for the multimodal mode-optimized group than when the treatment received was mode-invariant multimodal ($t(1566) = 2.87$, $p = .004$), spoken mode-invariant ($t(1585) = 3.55$, $p < .001$), spoken mode-optimized ($t(1467) = 3.34$, $p < .001$), textual mode-invariant ($t(2517) = 3.80$, $p < .001$), or textual mode-optimized ($t(2471) = 3.29$, $p = .001$) definitions, as well as the control treatment with no definitions ($t(2471) = 8.99$, $p < .001$). Once again, the effectiveness of multimodal definitions is due to the use of complementary, rather than mode-invariant, instructions.

Respondent Burden

Survey respondents' acceptance of multimodal clarification, particularly compared to unimodal formats, is crucial if multimodal definitions are realistically to be deployed in production research. If respondents react negatively to

multimodal communication, potentially abandoning the survey, these perceptions must be weighed against the increase in data quality brought about by this approach to clarification in online surveys, at least as demonstrated here.

To explore this, we asked respondents to rate their satisfaction with the survey and how burdensome they found the process; we also measured the amount of time respondents spent on each page of the web survey. The number of seconds respondents spent on the 12 survey items with definitions is shown in Table 6. Comparing mean response times with a Tukey adjustment, respondents with spoken mode-invariant definitions spent significantly more time completing the questionnaire than spoken mode-optimized or textual respondents. Respondents with multimodal definitions spent significantly more time than textual mode-invariant respondents, but not significantly longer than other types of definitions.

Table 6 Time spent on 12 definition questions by definition mode and optimization (in seconds)

Definition mode	25th percentile	Median	75th percentile	Mean	*SD*
Control (no definition)	72	93	136	107	51
Spoken mode-invariant	142	288	410	299	205
Spoken mode-optimized	133	198	279	237	175
Textual mode-invariant	90	130	209	169	120
Textual mode-optimized	105	160	195	192	219
Multimodal mode-invariant	140	222	363	264	171
Multimodal mode-optimized	142	192	281	249	235

However, a longer survey duration does not necessarily indicate that respondents *feel* more burdened. Respondents who were presented with unimodal spoken and multimodal definitions were asked to describe how burdensome they found the process of accessing spoken definitions (Not at all burdensome, slightly burdensome, moderately burdensome, very burdensome, extremely burdensome). This question was designed to measure the effort required to play spoken definitions, and so is not applicable to respondents in the control group, who saw no definitions, or respondents who were assigned to view unimodal textual definitions, since textual definitions appeared by default with no additional action needed from respondents. Overall, respondents did not indicate that playing definitions was notably burdensome. Most reported that accessing definitions was not at all burdensome (61%) or slightly burdensome (21%), while few found the process to be very (4%) or extremely (3%) burdensome. Multimodal respondents had the option of reading definitions without deliberately playing spoken definitions, so it is notable that the perceived level of burden did not vary between these four

types of definitions ($\chi^2(4) = 1.33$, $p = .856$). That is, respondents found the process of playing definitions to impose little burden regardless of whether they had another option for obtaining that information.

We also asked respondents to rate their overall satisfaction with the survey (Overall, how satisfied were you with your experience when responding to this survey? Very dissatisfied, somewhat dissatisfied, neither dissatisfied nor satisfied, somewhat satisfied, very satisfied). Respondents provided positive feedback about their survey experience. Almost half (48%) were very satisfied, and one-third (33%) were somewhat satisfied. The remainder were neither dissatisfied nor satisfied (14%), somewhat dissatisfied (4%), or very dissatisfied (1%). This distribution differed by definition mode ($\chi^2(12) = 32.28$, $p < .001$), with relatively higher proportions of respondents who were exposed to unimodal spoken and multimodal definitions reporting they were very satisfied when compared to unimodal textual respondents and those who were not shown definitions (53%, 52%, 39%, and 31%, respectively). Together, these suggest that multimodal definitions can be implemented in online surveys without overburdening respondents or otherwise causing a negative survey experience.

Discussion

Why were multimodal definitions—especially when optimized—more effective than unimodal definitions? On the one hand, if speech and written text are processed at least somewhat independently, then any multimodal communication (fully redundant or complementary) would improve comprehension when compared to unimodal communication, since more information would be available to a respondent, potentially compensating for an attentional lapse and providing more opportunity to internalize the content. Alternatively, if redundant definitions are less effective than complementary (i.e., mode-optimized) multimodal definitions at conveying the intended meaning of the question, then the latter should improve response quality more than unimodal definitions. Responses based on multimodal definitions were more aligned with survey concepts than responses based on unimodal definitions, and this was driven by mode-optimized definitions. This suggests that it is primarily complementary, rather than redundant multimodal content, that is effective (and supporting the idea that conveying identical information through multiple channels can reduce—or at least not facilitate—comprehension). The increased alignment with multimodal, and particularly mode-optimized multimodal definitions, appeared when comparing all observations. While the presence of multimodal definitions (regardless of whether they were used) increased data quality, these cues alone did not prompt respondents to attend to definitions; instead, the effect of multimodal definitions was sharpened when analyses were restricted to all compliant obser-

vations, as compliance with instructions about how to use definitions provided a purer measure of their impact on comprehension.

Overall, compliance was higher for mode-optimized than for mode-invariant definitions. Because the features of optimization (e.g., concision, increased salience of key material) were presented as a package and not experimentally varied, we cannot determine which of these features may have been most responsible for its benefits in multimodal definitions. In fact, for all definition types, compliance was highest for the first survey item than for subsequent questions, but respondents were willing to play spoken definitions in a survey mode that typically includes only text. While compliance could perhaps increase with shorter or more visually appealing definitions (two features that differentiated mode-invariant and mode-optimized presentations), these findings are promising for the efficacy of multimodal definitions, particularly given the strict compliance criteria for spoken definitions (i.e., respondents were required to fully play an audio clip). If respondents only minimally complied with multimodal definitions, or if they provided negative feedback about their experiences, those drawbacks would have to be carefully weighed against the increased alignment with definitions for responses to multimodal instructions. Instead, these results suggest that respondents do not find multimodal definitions to be burdensome, are willing to comply with instructions to both read and listen to them, and will apply these definitions to their formulation of survey responses. In an online survey, multimodal definitions can improve data quality without negatively impacting respondents. It is reassuring that the presence of spoken information did not decrease respondent satisfaction, and in fact, respondents who were presented with spoken definitions either alone or as part of multimodal definitions reported the highest levels of satisfaction.

Future Research

The sample for this study was drawn from Amazon Mechanical Turk. This study provides a proof-of-concept that multimodal definitions can improve data quality, but more research is needed to determine the degree to which these findings can be replicated in samples from other sources and whether unpaid participants are as amenable to play integrated audio clips in an online survey. We were unable to capture the type of device on which surveys were completed, for example, a laptop computer or smartphone, and these findings may vary further by device type.

Compliance was inferred without truly knowing whether respondents attended to definitions. For spoken definitions, compliance may have been underestimated for respondents who partially listened to spoken definitions. For textual definitions, compliance may have been over- or under-estimated if

respondent reading speed was miscalculated by our use of response latency as a measure, or if they simply did not attend to their screen. For spoken definitions, a more robust tracking mechanism could assess how much of spoken definitions were played. For textual definitions, a lab study that tracks respondents' eye movements could more accurately measure whether on-screen text was read. All of these limitations can be addressed in straightforward ways in follow-up studies.

This study focuses on a fundamentally visual type of survey: a textual web survey, in which spoken definitions were embedded in some experimental conditions. While text is persistent, spoken communication is ephemeral, so improvements in data quality due to adding text to a communication format that is typically spoken (such as telephone surveys) is likely to be greater than the improvements due to adding spoken information to a communication format that is typically textual (such as web surveys). While some spoken surveys do have an added textual component (e.g., show cards), that text has typically been used to present response options, rather than questions and definitions. Telephone surveys rarely include a textual component, and this gap is particularly ripe for exploration. Respondents completing a telephone survey are often using an internet-enabled device. A respondent could receive text instructions from an interviewer, e.g., via a text message, particularly for survey items for which the underlying constructs are nuanced or potentially counterintuitive. While the effectiveness of multimodal communication may differ across these scenarios, particularly given differences in communication norms and respondent expectations, these uses warrant further exploration of multimodal definitions given its richness, the likelihood it will become more practical with technological advances, and the possibility that respondents will be more satisfied with their experience knowing they understand what they are asked.

References

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading, 36*(2), 84–95. http://www.jstor.org/stable/40016440

Clark, H. H. (1996). Community, commonalities, and communication. In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 324–355). Cambridge University Press.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods, 11*(1), 45–61. https://doi.org/10.18148/srm/2017.v11i1.6304

Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64*(1), 1–28. https://doi.org/10.1086/316757

Conrad, F. G., & Schober, M. F. (2021). Clarifying question meaning in standardized interviews can improve data quality even though wording may change: A review of the evidence. *International Journal of Social Research Methodology, 24*(2), 203–226. https://doi.org/10.1080/13645579.2020.1824627

Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing features of human dialogue to web surveys. *Applied Cognitive Psychology, 21*(2), 165–187. https://doi.org/10.1002/acp.1335

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. In D. Lalanne & J. Kohlas (Eds.), *Human machine interaction* (pp. 3–26). Springer. https://doi.org/10.1007/978-3-642-00437-7_1

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. https://doi.org/10.1037/h0057532.

Johnston, M. (2008). Automating the survey interview with dynamic multimodal interfaces. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the survey interview of the future*. Wiley. https://doi.org/10.1002/9780470183373

Kalyuga, S., Chandler, P., & Sweller, J. (2004). When redundant on-screen text in multimedia technical instruction can interfere with learning. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(3), 567–581. https://doi.org/10.1518/hfes.46.3.567.50405

Leahy, W., & Sweller, J. (2011). Cognitive load theory, modality of presentation and the transient information effect. *Applied Cognitive Psychology, 25*(6), 943–951. https://doi.org/10.1002/acp.1787

Mayer, R. E. (Ed.). (2014). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning* (2nd ed., pp. 72–103). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*(1), 187–198. https://doi.org/10.1037/0022-0663.93.1.187

Mayer, R. E., & Johnson, C. I. (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology, 100*(2), 380–386. https://doi.org/10.1037/0022-0663.100.2.380

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*(1), 156–163. https://doi.org/10.1037/0022-0663.94.1.156

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*(2), 319–334. https://doi.org/10.1037/0022-0663.87.2.319

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *The Public Opinion Quarterly, 61*(4), 576–602. https://doi.org/10.1086/297818

Schober, M. F., Conrad, F. G., & Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology, 18*(2), 169–188. https://doi.org/10.1002/acp.955

Schober, M. F., Suessbrick, A. L., & Conrad, F. G. (2018). When do misunderstandings matter? Evidence from survey interviews about smoking. *Topics in Cognitive Science, 10*(2), 452–484. https://doi.org/10.1111/tops.12330

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology: General, 119*(2), 176–192. https://doi.org/10.1037/0096-3445.119.2.176.

Sweller, J., Van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*, 261–292. https://doi.org/10.1007/s10648-019-09465-5

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018). Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society Series A: Statistics in Society, 181*(1), 181–203. https://doi.org/10.1111/rssa.12255

White, S. (2012). Mining the text: 34 text features that can ease or obstruct text comprehension and use. *Literacy Research and Instruction, 51*, 143–164. https://doi.org/10.1080/19388071.2011.553023

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods, 8*, 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453

Appendix

Definitions by Survey Question and Treatment Group

Question	In the past 7 days, how many hours of television did you watch?
Mode-invariant definition	Watching television includes any programs other than films. This may include sitcoms, dramas, news, sports, and reality shows. Television is watched on a television set at the time it is broadcast and does not include programming recorded with a DVR, viewed on-demand, or streamed. Include content viewed on a television set only and exclude any content viewed on a computer or mobile device.
Textual mode-optimized definition	• **Content is broadcast.** Exclude DVRed, on-demand, and streamed shows. • **TV set.** Exclude shows watched on a computer or mobile device. • **TV shows.** Exclude films, even if watched while they air.
Spoken mode-optimized definition	By television, we mean content watched on a TV set at the time it is broadcast. Exclude streamed, on –demand, and DVRed shows and anything watched on a computer or mobile device. Exclude films.
Inclusive/exclusive	Exclusive definition

Question	In the past 7 days, for how many hours did you listen to the radio?
Mode-invariant definition	Listening to the radio includes listening to programming transmitted and received through an antenna. Available stations and reception are restricted by signal strength and listener location. Programs are listened to live, that is, as they air, rather than played later by the listener, such as with podcasts and other downloadable content. Programming can include talk-based content, such as news or sports, but does not include music even if accessed by antenna.
Textual mode-optimized definition	• **Antenna.** Only count local stations through over-the-air access, not satellite or internet. • **Live Content.** Exclude podcasts or other content played on-demand. • **Talk.** Programming includes news, sports, and talk shows. Exclude music.
Spoken mode-optimized definition	By radio, we mean local programming listened to live, over-the-air, but not podcasts, satellite radio, or internet radio. Include news, sports, and talk shows, not music.
Inclusive/exclusive	Exclusive definition

Question	In the past 7 days, for how many hours did you use e-mail?
Mode-invariant definition	E-mail use includes composing, sending, and reading messages, as well as managing an inbox. Count time spent using an online mailbox, desktop mailbox, or mobile application, and do not count time spent reading attachments or linked content in a browser. Only count e-mail use when connected to the internet through a wired or wireless (Wi-Fi) connection. Exclude email use involving a cellular connection such as 3G or 4G. Exclude offline use.
Textual mode-optimized definition	**Exclude** • E-mail using a cellular network such as 3G or 4G. • Reading attachments or linked content. **Include** • Composing, sending, reading, and sorting messages. • Use of a Wi-fi or wired connection.
Spoken mode-optimized definition	By e-mail use, we mean writing, reading, sending and sorting messages. Only count time using an application, not time spent reading attachments or linked content. Only count access through a wired or Wi-Fi connection, so exclude cellular networks such as 3G and 4G.
Inclusive/exclusive	Exclusive definition

Question	Excluding e-mail use, in the past 7 days, for how many hours did you use the internet?
Mode-invariant definition	People may use the Internet to carry out personal or professional tasks and activities. Exclude internet use involving a cellular connection such as 3G or 4G. Include active tasks such as reading news articles, posting in online forums, and playing online games. Exclude passive tasks that do not involve direct attention or engagement such as streaming videos or music.
Textual mode-optimized definition	• **Connection.** Count Wi-fi and wired connections only. Exclude cellular networks such as 3G and 4G. • **Active use.** Count tasks such reading articles, posting in forums, and playing online games. Do not count passive activities such as streaming videos or music.
Spoken mode-optimized definition	By Internet, we mean access through a wired or Wi-Fi connection, so exclude cellular networks such as 3G and 4G. Only count time on tasks such as reading or posting content or playing games, and do not count passive activities such as streaming videos or music.
Inclusive/exclusive	Exclusive definition

Question	In the past 7 days, how many hours did you work in total?
Mode-invariant definition	Work is paid employment performed for an employer or, if self-employed, for oneself. Count paid internships or apprenticeships. Count time directly spent on work activities, such as time at an office or work site, as well as commuting to and from an office.
Textual mode-optimized definition	**Include** • Paid work or self-employment. • Work as an employee or paid intern. • Time at work and **commuting** to and from work.
Spoken mode-optimized definition	By work, we mean a paid job or internship, or self-employment. In addition to time at a job site, work includes commuting time.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many miles did you travel by vehicle?
Mode-invariant definition	Vehicles have two or more wheels, are used for ground transportation and can include cars, trucks, taxis, buses, trains, subways, trams, motorcycles, and bicycles. All miles spent in a vehicle, regardless of seat location, should be considered. Miles as both a driver and passenger should be included.
Textual mode-optimized definition	• **Vehicle.** Count any ground travel by vehicle, including cars, trucks, taxis, buses, motorcycles, trains, subways, and bicycles. • **Role.** Count miles as both driver and passenger.
Spoken mode-optimized definition	By travel, we mean miles as a driver or passenger in a vehicle such as a car, truck, taxi, bus, train, subway, tram, motorcycle, or bicycle.
Inclusive/exclusive	Inclusive definition

Question	In the past year, how many plane trips did you take?
Mode-invariant definition	A plane trip begins at liftoff and ends at touchdown. If multiple legs (liftoffs and touchdowns) are involved, such as with non-direct or multi-city flights, each is counted separately. Similarly, for roundtrip flights, outbound and return flights are each counted separately, and all legs are counted separately.
Textual mode-optimized definition	• Count each leg of a trip separately. • Count roundtrip flights separately.
Spoken mode-optimized definition	Count each component of a trip separately. For example, layovers and roundtrip flights should be counted as multiple plane trips.
Inclusive/exclusive	Inclusive definition

Question	In the past 30 days, how many times have you had food or drinks at a restaurant?
Mode-invariant definition	Restaurants are dining establishments at which food and/or beverages are served. Include sit-down establishments, restaurants with and without table service, fast food restaurants, coffee shops and cafes, bars and pubs, food trucks, and street vendors. Food may be eaten at the restaurant or elsewhere, if ordered for take-out, to-go, or delivery.
Textual mode-optimized definition	• **Type.** Count sit-down restaurants, fast food, coffee shops, bars, food trucks and street vendors. • **Location.** Count dine-in, take-out, to-go orders, and delivery.
Spoken mode-optimized definition	We mean sit-down restaurants, fast food, coffee shops, bars, food trucks and street vendors. We mean dine-in, take-out, to-go orders, and delivery.
Inclusive/exclusive	Inclusive definition

Question	How many pairs of shoes do you own?
Mode-invariant definition	Shoes are footwear worn primarily outdoors and secured to a foot with some type of fastener, such as laces, zipper, Velcro, clasps, or buckles. For this question, footwear designed primarily for indoor use such as slippers does not qualify. For this question, non-fastening shoes such as flip flops, slides, clogs, pumps, and other unsecured footwear do not qualify.
Textual mode-optimized definition	**Exclude** shoes • Worn indoors, including slippers. • Unsecured, such as flip flops, slides, clogs, pumps, etc. **Include** shoes • Worn outside • Secured with laces, zippers, Velcro, clasps, buckles, etc.
Spoken mode-optimized definition	By shoes, we mean footwear worn primarily outside that can be secured with fasteners such as laces, zippers, Velcro, clasps, or buckles. Do not count unsecured footwear such as flip flops, slides, clogs, pumps, and other unsecured footwear.
Inclusive/exclusive	Exclusive definition

Question	How many hours of rest do you get on a typical weekday?
Mode-invariant definition	Include time spent in a state of sleep or time that has the potential to become sleep. This includes overnight sleep and daytime naps, as well as time when sleep is not necessarily intended, such as during class or a meeting, while reading a book, or while watching television.
Textual mode-optimized definition	• **Time of day.** Count evening and daytime rest. • **Sleep state**. Count time spent asleep or when sleep is possible, such as sitting while reading a book or watching television.
Spoken mode-optimized definition	By rest, we mean time when you are asleep or could fall asleep, such as sitting while reading a book or watching TV.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many hours did you exercise?
Mode-invariant definition	Exercise is physical activity that results in an elevated heart rate. This can include vigorous activities such as running or biking and less vigorous activities such as walking, climbing up or down stairs, and yoga. Exercise can be performed alone, such as swimming or biking, or with a group or team, such as basketball or tennis. Include all physical activities, regardless of how long they lasted.
Textual mode-optimized definition	• **Activities.** Count all activities that result in an elevated heart rate. • **Duration.** Count all physical activities, regardless of how long they lasted.
Spoken mode-optimized definition	By exercise, we mean activities that result in an elevated heart rate, regardless of the duration of each activity.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many caffeinated drinks did you have?
Mode-invariant definition	Caffeine is a stimulant often found in cacao plants and a variety of beverages. Common caffeinated beverages include coffee, tea, and sodas. While caffeinated beverages may be consumed in any amount or container size, for this question, 8 fluid ounces of a caffeinated beverage is one caffeinated drink.
Textual mode-optimized definition	• Count every 8 ounces as one drink. • Count coffee, tea, soda, and other caffeinated beverages.
Spoken mode-optimized definition	By caffeinated drinks, we mean 8 ounces of caffeinated beverages such as coffee, tea, and soda.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many hours did you exercise?
Mode-invariant definition	Exercise is physical activity that results in an elevated heart rate. This can include vigorous activities such as running or biking and less vigorous activities such as walking, climbing up or down stairs, and yoga. Exercise can be performed alone, such as swimming or biking, or with a group or team, such as basketball or tennis. Include all physical activities, regardless of how long they lasted.
Textual mode-optimized definition	• **Activities.** Count all activities that result in an elevated heart rate. • **Duration.** Count all physical activities, regardless of how long they lasted.
Spoken mode-optimized definition	By exercise, we mean activities that result in an elevated heart rate, regardless of the duration of each activity.
Inclusive/exclusive	Inclusive definition

Question	In the past 7 days, how many caffeinated drinks did you have?
Mode-invariant definition	Caffeine is a stimulant often found in cacao plants and a variety of beverages. Common caffeinated beverages include coffee, tea, and sodas. While caffeinated beverages may be consumed in any amount or container size, for this question, 8 fluid ounces of a caffeinated beverage is one caffeinated drink.
Textual mode-optimized definition	• Count every 8 ounces as one drink. • Count coffee, tea, soda, and other caffeinated beverages.
Spoken mode-optimized definition	By caffeinated drinks, we mean 8 ounces of caffeinated beverages such as coffee, tea, and soda.
Inclusive/exclusive	Inclusive definition