

Survey item placement has little effect on data quality or treatment-effect estimates

Joohye Jeong^{1,*}, Adam J. Berinsky¹, John V. Kane² & Ning Soong¹

¹Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, USA. ²Center for Global Affairs, New York University, New York, NY, USA.

*e-mail: joohye@mit.edu

Abstract

Low data quality threatens the scientific merit of survey-based research. Researchers routinely screen out inattentive respondents at the start of a survey, but this does not guarantee that respondents remain attentive throughout. Researchers contributing items to omnibus surveys or deciding where to place key components, such as experiments, therefore need to know whether an item's position within a survey affects data quality. Existing evidence, however, is limited and inconsistent. Here we report five high-powered surveys ($N = 15,159$) that randomly assigned an extensive range of data-quality measures, and (in three surveys) a diverse set of previously published survey experiments, to early, middle, or late positions. Position effects on data-quality measures were generally modest, mixed in direction and inconsistent across item types. Treatment-effect estimates remained stable across positions. Taken together, these findings indicate that item and experiment placement within a survey is unlikely to matter substantially for data quality or estimated treatment effects.

19 Surveys are among the most widely used instruments for studying human attitudes
20 and behaviour. Yet, the validity of many survey measures depends on respondents' effortful
21 reading and consideration of the questions presented to them. When respondents engage
22 carelessly with those questions, the consequences can be substantial. Careless responding
23 can distort estimates of population means and correlations between items, and attenuate
24 estimated treatment effects in survey experiments¹⁻³.

25 Researchers commonly use attention checks to "screen out" low-effort respondents
26 before the survey begins, but passing such a screener does not ensure that a respondent
27 remains sufficiently attentive for the duration of the study⁴. Potential variation in attentive-
28 ness raises concerns for all survey research, but the concern may be especially pronounced
29 in multi-investigator omnibus surveys, which are often relatively long and contain several
30 unrelated modules or experiments. Researchers facing resource constraints may include sev-
31 eral experiments within a single study, and separate research teams routinely submit items
32 to one large shared survey, as in cooperative multi-investigator studies⁵. Although these
33 practices use resources efficiently, the longer surveys they produce raise important practical
34 questions. To what extent does data quality—as measured by, for example, attention to
35 particular items, response times or willingness to answer open-ended questions—change as
36 a survey progresses? Are survey components such as individual questions, question grids
37 or vignette-based experiments systematically disadvantaged when placed at the beginning,
38 middle, or end of a survey?

39 Despite the widespread reliance on lengthy, omnibus surveys, researchers currently
40 lack consistent, robust evidence on these critical questions. On one hand, several stud-
41 ies report significant variability in data quality over the course of a survey. Gibson and
42 Bowling⁶, for example, find some support for the hypothesis that survey length affects the
43 likelihood of providing correct and consistent responses, as do Bowling et al.⁷, although
44 the latter results depend somewhat on the particular measure of data quality employed.
45 Galesic and Bosnjak⁸ find that questions placed later in a survey tend to be answered more

quickly and, in the case of question grids, with less variability, a behaviour indicative of “straightlining”. On the other hand, a recent analysis of YouGov panel data finds that respondents were significantly more likely to pass the same attention check when it appeared towards the end of a survey rather than at the start (<https://yougov.com/en-us/articles/52081-attention-check-placement-and-formatting>).

However, other research finds limited evidence of variability over the course of a survey. Bäckström and Björklund⁹ report that response times were faster in the second half of a survey but that the data were no less reliable. Similarly, Bansak et al.¹⁰ find that employing more tasks in conjoint experiments, which lengthens the study, does not reduce data quality. And although attentiveness could plausibly change in response to the inclusion of attention checks themselves, there is no evidence for this hypothesis either^{11–13}.

The evidence, in short, does not clearly point to either waxing or waning of data quality over the course of a long survey. Further, the existing evidence is limited in several respects. First, some studies rely on student or otherwise non-representative samples⁶, making it difficult to assess the variability of data quality in the large national samples that social scientists typically seek. Second, existing work has tended to employ a narrow set of data-quality measures, which matters given the many ways in which data quality can be both defined and assessed¹⁴. Recent work has documented substantial variation in data quality *across* items (as well as across platforms and samples)¹⁵, but the more fundamental question of whether data quality significantly varies *within* a survey (i.e., across positions in the questionnaire) remains unresolved. Third, previous designs rarely placed data-quality measures far apart within the survey¹⁰, nor did they use enough survey positions to detect *non-monotonic* variation in data quality⁹. This is an important omission given the plausibility of non-monotonic variation: at the start of a survey, for example, respondents may remain distracted by whatever they were engaged in immediately prior to the survey’s start, then slowly “warm up” as the survey progresses, but later experience some amount of fatigue and respond more carelessly⁸.

A final limitation of existing work is that researchers have not examined whether experimental treatment effects systematically vary over the course of a survey. To the extent that data quality varies, we should also expect experimental treatment effects to vary¹. Yet stable (or even improving) performance on brief attention checks does not imply that respondents will carefully read experimental vignettes or respond effortfully to outcome measures¹⁶. This matters because inattentiveness will typically attenuate treatment effect estimates¹⁷. Whether treatment effects vary according to their position in a survey, therefore, is a second critically important question for survey researchers.

Thus, given the limited and mixed evidence in the existing literature, our study tests two primary hypotheses: that data quality depends on the position of the measure within the survey (H1), and that experimental treatment effects depend on the position of the experiment (H2). Both hypotheses are non-directional, as data quality and treatment effects could plausibly increase (a learning or “warm-up” effect), decrease (a fatigue effect) or follow a non-monotonic pattern.

We examine these questions in five large surveys fielded on Lucid Theorem between February 2020 and August 2025 (total $N = 15,159$ completed responses; Table 1). Lucid is a widely used, low-cost online platform that provides quota-sampled panels approximating US Census demographics¹⁸. Because Lucid samples have exhibited lower average response quality than samples from several other online platforms¹⁵, they provide an informative setting in which to test whether later placement compounds existing data-quality concerns.

Each survey randomly assigned measures to positions near the beginning (Position 1), approximately halfway through (Position 2) or near the end (Position 3) of the questionnaire. Crucially, this allows for the testing of *non-monotonic* position effects. The five surveys span complementary designs, from fully within-subjects (each respondent encountering measures at multiple positions) to pure between-subjects (each respondent receiving a single measure at a single position; Survey 4), which separates position effects from within-survey learning. The analysis plan was registered while Survey 5 was in the field, before its data were ex-

amined, following exploratory analyses of Surveys 1–4; further details are provided in the Methods.

We assessed data quality with seven distinct types of measures (Table 2): instructed-response items (IRIs), which embed a directive that respondents must follow¹; bogus items, which present implausible or verifiable statements; factual manipulation checks (FMCs), which test recall of experimental content¹⁹; mock vignette checks (MVCs), which test comprehension of practice vignettes¹⁷; and three behavioural indicators, namely vignette reading time, straightlining across grid items and the substantiveness of open-ended responses. Because these measures differ in format, difficulty and the aspect of engagement they capture, we can assess whether position effects generalize across distinct components of data quality (full item wordings in Supplementary Note 2). We refer to this full set collectively as *data-quality measures*, recognizing that they capture distinct aspects of respondent engagement rather than a single latent construct.

Table 1: Summary of the five surveys

	<i>N</i>	Field date	Positions	Avg./med. duration (min.)	Data-quality measures	Experiments
Survey 1	5,030	Feb 2020	1,2	27.8 / 18.3	MVCs ^M , FMCs, Reading time	Student loan, KKK, Welfare, Immigration
Survey 2	2,013	Nov 2024	1,3	28.7 / 15.6	IRIs (Simple, Web, Color ^R)	—
Survey 3	3,024	Mar–Apr 2025	1–3	28.3 / 16.6	IRIs (Simple, Web, Color), FMCs, Reading time	Student loan, Welfare, Disease
Survey 4	1,414	Jul 2025	1–3	38.9 / 15.5	IRIs (Web, Color) ^B	—
Survey 5	3,678	Jul–Aug 2025	1–3	22.3 / 14.7	IRIs (Section, Issue), Bogus (Brush, Grid), FMCs, Straightlining, Substantive open-ended response, Reading time	Student loan, Welfare, Disease (short/long)

Note: ^MAt Position 1, 20.5% of Survey 1 respondents were randomly assigned not to receive an MVC module, whereas all were assigned an MVC module at Position 2; respondents receiving modules at both positions encountered different vignettes. ^RColor item repeated (presented twice). ^BBetween-subjects: each respondent received one item at one position. IRI = instructed-response items; Bogus = implausible or verifiable statements; FMCs = factual manipulation checks; MVCs = mock vignette checks.

Table 2: Data-quality measures

Category	Item	Description	Surveys
<i>Performance-based measures</i>			
IRIs	Simple	Follow a directive to select a specific response option	2, 3
	Web	Select two named outlets in a news-website prompt	2, 3, 4
	Color	Select two named colours in a favourite-colour prompt	2, 3, 4
	Section	Select two named sections in a newspaper-reading prompt	5
	Issue	Rank two named issues in specified positions	5
Bogus items	Brush	Implausible self-report: “I have never brushed my teeth”	5
	Grid ($\times 2$)	Verifiable facts in grid format (for example, “Two is greater than one”)	5
FMCs	—	Factual recall of experimental vignette content	1, 3, 5
MVCs	—	Comprehension of practice (non-treatment) vignettes	1
<i>Behavioural indicators</i>			
	Straightlining	Identical responses across a 7-item opinion grid	5
	Substantive open-ended response	Response to open-ended probe; ≥ 5 characters, containing ≥ 1 English word	5
	Reading time	Seconds spent on mock and experimental vignette page	1, 3, 5

Note: IRI = instructed-response items (including screeners and IMCs); Bogus = implausible or verifiable statements; FMCs = factual manipulation checks; MVCs = mock vignette checks. Full item wordings are provided in Supplementary Note 2.

Three surveys additionally randomized the placement of five published survey experiments, each of which has been successfully replicated in prior work: a welfare-framing experiment²⁰, the classic disease-risk framing problem²¹, a student-loan framing experiment²², a KKK rally experiment²³ and an immigration experiment²⁴. The latter two appeared only in Survey 1. To facilitate comparison and pooling, we rescaled all outcomes to a common 0–1 metric and coded their direction so that positive values corresponded to the treatment direction defined in the original studies (see Methods for full details of the samples, measures,

designs and analyses).

Across the five surveys, we found limited evidence that data quality or treatment-effect estimates depended systematically on position. Position differences on the data-quality measures were generally modest and varied in direction, the largest being an approximately 6-percentage-point increase in IRI pass rates at later positions; treatment-effect estimates were stable, with pooled position contrasts close to zero. Equivalence tests bounded all but one contrast within a conventional small-effect benchmark²⁵, indicating small, precisely estimated differences rather than uninformative null results.

Results

Data quality by survey position

To the extent that respondents become fatigued over the course of a survey, we should observe that respondents perform worse on data quality measures as a survey progresses. Alternatively, if respondents incrementally become more comfortable with the process of reading and answering questions (i.e., “warm-up”) we should observe improvement on these measures. Finally, if *both* processes are occurring at different points in the survey, we should still see systematic variability across positions.

Figure 1 shows adjusted mean performance on each data-quality measure at each position. Because not every survey fielded every measure at every position (Table 1), these profiles average model predictions (predictive margins) over the same survey–item composition at every position, so that movement along a line reflects position rather than a changing mix of surveys and items. Formal pairwise comparisons instead use a two-stage stratified estimator, which forms each contrast only within survey–item strata observed at both positions and combines them with sample-size weights; the estimates therefore need not equal the simple differences between the plotted means (see Methods). Unless noted otherwise, all reported contrasts, P values and equivalence tests use this estimator.

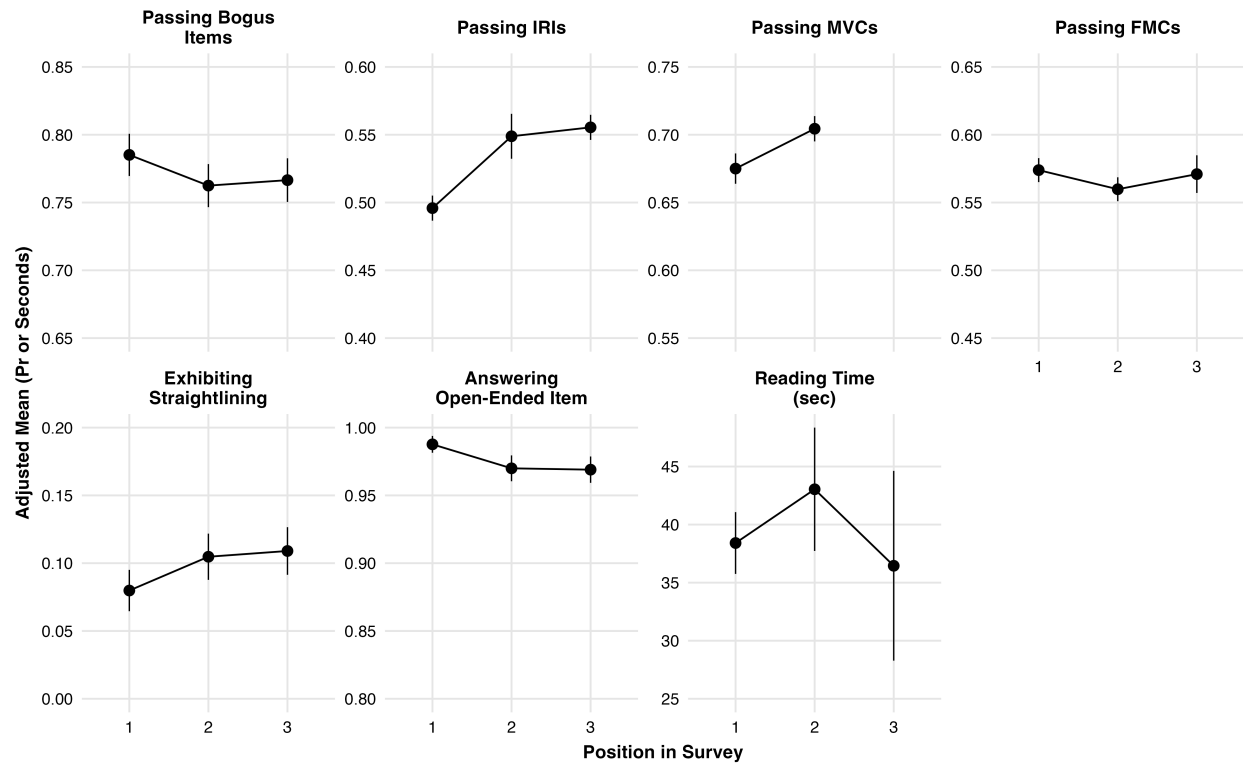


Figure 1: **Data quality by survey position.** Model-based adjusted means (predictive margins) for bogus-item, IRI, MVC and FMC pass rates, straightlining, substantive open-ended responses and vignette reading time (total $n = 15,159$). Estimates are from linear models with survey-item cell fixed effects and additive position interactions by survey and item, averaged over the same empirical survey-item composition at every position. Error bars show 95% confidence intervals based on respondent-clustered standard errors. Formal pairwise contrasts are reported in Supplementary Table 3, and an observed-support version of this figure in Supplementary Fig. 1.

Position differences were generally modest and varied in direction across measures, providing no evidence of a uniform deterioration or improvement in data quality. Performance on some measures improved modestly at the middle and later positions. IRIs showed the clearest pattern: adjusted pass rates were 49.6% at Position 1, 54.9% at Position 2 and 55.5% at Position 3. The corresponding contrasts were +5.8 percentage points for Position 2 versus 1 (95% confidence interval (CI) 4.4 to 7.2) and +6.0 percentage points for Position 3 versus 1 (95% CI 4.9 to 7.0; both $P < 0.001$).

MVCs, measured only in Survey 1, showed an aggregate increase from Position 1 to 2 (+2.9 percentage points, 95% CI 1.9 to 3.9, $P < 0.001$). Survey 1's randomized Position 1

module assignment (Table 1) permitted an additional analysis by prior MVC-module exposure. Among first MVC-module exposures, the pass rate was nearly identical at Position 1 (67.5%) and Position 2 (67.4%; -0.1 percentage points, 95% CI -2.5 to 2.3 , $P = 0.940$). At Position 2, the pass rate was 71.2% among respondents who had completed a different MVC module at Position 1 and 67.4% among those encountering their first MVC module ($+3.8$ percentage points, 95% CI 1.4 to 6.2 , $P = 0.002$). Thus, the aggregate improvement was concentrated among respondents with prior MVC-format exposure, suggesting that the better performance on the latter item is not due to its position *per se*, but to respondents learning to expect follow-up questions after a vignette.

Other measures moved in the opposite direction or changed little. Adjusted bogus-item pass rates declined from 78.5% at Position 1 to 76.3% and 76.7% at Positions 2 and 3. The corresponding contrasts were -2.3 percentage points for Position 2 versus 1 (95% CI -4.4 to -0.1 , $P = 0.038$) and -1.9 percentage points for Position 3 versus 1 (95% CI -4.0 to 0.3 , $P = 0.088$). FMC pass rates also declined slightly: -1.4 percentage points for Position 2 versus 1 (95% CI -2.6 to -0.3 , $P = 0.016$) and -1.4 percentage points for Position 3 versus 1 (95% CI -2.9 to 0.2 , $P = 0.085$). Straightlining increased by 2.5–2.9 percentage points at later positions, while substantive open-ended responses declined by 1.8–1.9 percentage points from a near-ceiling baseline of 98.8%. Vignette reading-time contrasts were not statistically distinguishable from zero ($P = 0.123$ for Position 2 versus 1 and $P = 0.883$ for Position 3 versus 1).

The 13 pairwise contrasts all bear on the same hypothesis, that data quality depends on survey position (H1), and testing this many contrasts individually risks mistaking chance fluctuations for position effects. As registered, we therefore adjusted for multiple testing, using the Benjamini–Hochberg (BH) procedure to control the false discovery rate across the full set of 13 contrasts. Both IRI contrasts, the MVC contrast, the Position-2 FMC contrast, both open-ended contrasts and the Position-3 straightlining contrast remained significant after adjustment. The bogus-item contrasts, Position-3 FMC contrast, Position-2

straightlining contrast and reading-time contrasts did not (all adjusted $P > 0.05$).

Because failing to reject a null hypothesis is not by itself evidence of a null effect, we also conducted equivalence tests²⁶, which reverse the burden of proof. The null hypothesis is a difference of at least a given size, so rejecting it provides positive evidence that any difference is small. For each contrast, we computed an equivalence confidence interval (ECI), a symmetric interval $\pm\delta$ that we report by its margin δ : the smallest value such that all effects larger than δ in magnitude can be statistically ruled out at $\alpha = 0.05$. Smaller ECIs therefore indicate tighter bounds on possible position effects. Across the 13 contrasts, ECIs ranged from 2.4 to 7.0 percentage points for binary data-quality measures and from 2.0 to 9.5 s for vignette reading time. Using the conventional small-effect benchmark of $\pm 0.20\sigma_Y$ (ref. 25), all 13 contrasts were statistically equivalent after BH adjustment (Supplementary Table 3).

To simplify exposition, the aforementioned results pooled several distinct types of bogus items and IRIs. Figure 2 therefore disaggregates these particular results by both item and survey. Among bogus items, the Brush item remained high across positions (82.5–85.5%), whereas the two Grid items generally declined: Grid Q3 moved from 75.8% to 68.5% and 70.5%, and Grid Q5 from 77.2% to 74.7% and 74.0% (Fig. 2a).

Among IRIs, the aggregate improvement was driven largely by the Color item, whose adjusted pass rate rose from 57.2% at Position 1 to 70.5% and 70.6% at Positions 2 and 3 (Fig. 2b). Prior exposure to an embedded instruction may have made respondents more alert to similar items later in the survey, but repeated exposure was not necessary: in Survey 4’s pure between-subjects design, Color pass rates rose monotonically from 43.8% at Position 1 to 55.6% and 60.6% at Positions 2 and 3 (Fig. 2c). The Web item improved at later positions in Surveys 2 and 3 (41.7% to 43.6% between Positions 1 and 3 in Survey 2; 31.2% to 42.2% and 46.5% in Survey 3), whereas in Survey 4 it followed an inverted U, rising from 24.6% at Position 1 to 33.1% at Position 2 before returning to 24.8% at Position 3 (Fig. 2d). The remaining IRIs displayed no consistent tendency with respect to data quality across

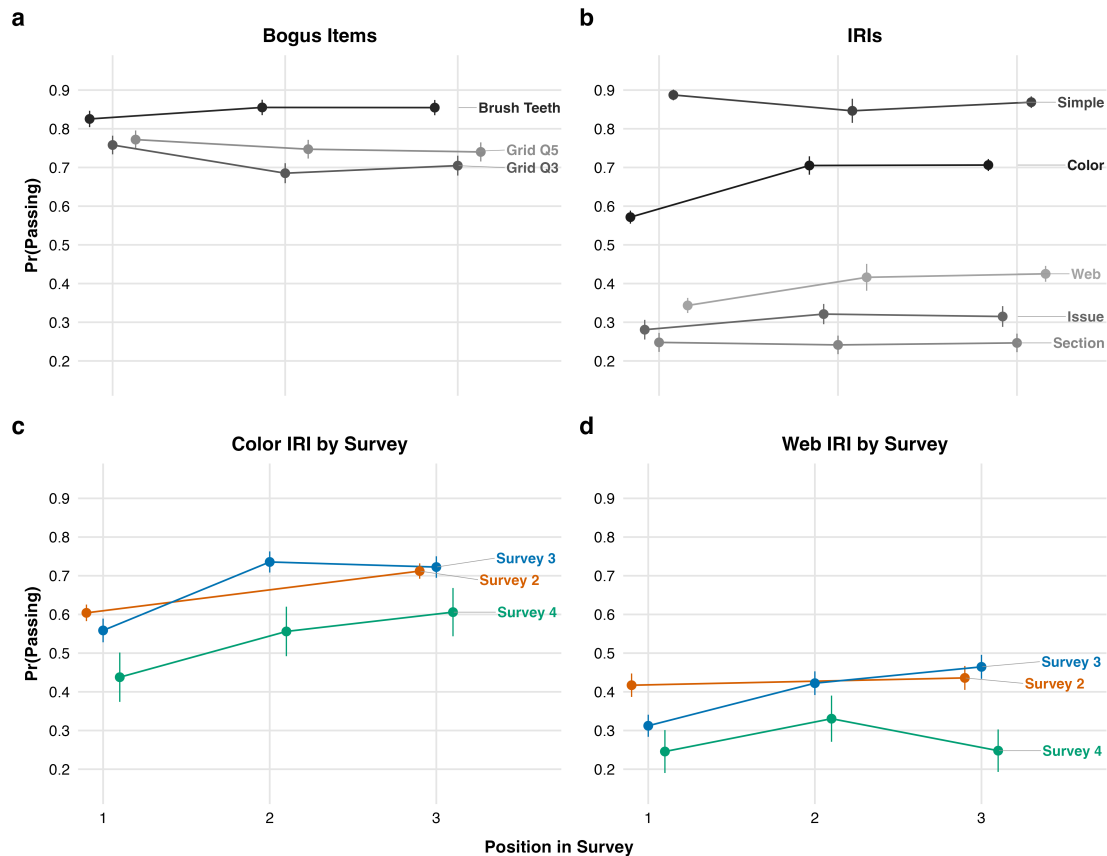


Figure 2: **Bogus-item and IRI pass rates by survey position.** **a**, Individual bogus items in Survey 5. **b**, Individual IRIs pooled across surveys; each line is the item's adjusted profile from the model used in Fig. 1, standardized over the item's full survey composition at every position. **c,d**, The Color and Web IRIs by survey; survey-specific lines show only positions observed in that survey. Survey 2 repeated the Color item, Survey 3 used a within-subjects design and Survey 4 used a between-subjects design. Values in **a**, **c** and **d** are observed cell means, which equal adjusted means for single-survey cells ($n = 2,013, 3,024, 1,414$ and $3,678$ in Surveys 2–5). Error bars show 95% confidence intervals based on respondent-clustered standard errors. An observed-support version of panel **b** is shown in Supplementary Fig. 2.

positions. The Simple item dipped modestly at Position 2, the Section item was nearly flat and the Issue item improved modestly. Together with the opposing bogus-item pattern, this heterogeneity suggests item- and design-specific processes rather than a general shift in data quality as the survey progresses.

Treatment effects by survey position

Although no consistent pattern of improving or worsening data quality emerged, some measures showed item-specific position effects. For survey experimentalists, however, the central question is whether the placement of an experiment changes the resulting treatment-effect estimate. If data quality (e.g., attentiveness) declines over the course of a survey, treatment effects may attenuate at later positions; conversely, if respondents warm up, effects may strengthen. Figure 3a,b shows adjusted treatment-effect profiles, pooled across all experiments and separately by each particular experiment. As in Fig. 1, these profiles are predictive margins averaged over a fixed mix of surveys and experiments at every position. Formal comparisons use the two-stage stratified estimator, here contrasting the estimated treatment effect across positions within each survey–experiment stratum; as with the data-quality measures, these contrasts need not equal the simple differences between the plotted points (see Methods).

The adjusted pooled profile was remarkably similar across positions (16.4%, 16.3% and 15.3% at Positions 1–3). The pooled treatment-effect differences were -0.1 percentage points for Position 2 versus 1 (95% CI -2.0 to 1.8 , $P = 0.915$), -1.1 percentage points for Position 3 versus 1 (95% CI -3.6 to 1.4 , $P = 0.399$) and $+0.1$ percentage points for Position 3 versus 2 (95% CI -2.5 to 2.7 , $P = 0.924$). The 3-versus-2 contrast is estimated on a different set of supported strata and therefore need not equal the difference between the plotted values.

Similarly, no individual experiment showed clear evidence of position-dependent treatment effects (Fig. 3b). We tested treatment-effect differences across positions within each of the five experiments, yielding 11 experiment-specific pairwise contrasts in addition to the three pooled contrasts. The smallest unadjusted P value was observed for the immigration experiment’s Position 2-versus-1 contrast ($+4.3$ percentage points, 95% CI -1.9 to 10.4 , $P = 0.172$); every other experiment-specific contrast had an unadjusted $P \geq 0.45$. After BH adjustment across all 14 treatment-effect contrasts, every adjusted P value was 0.993 .

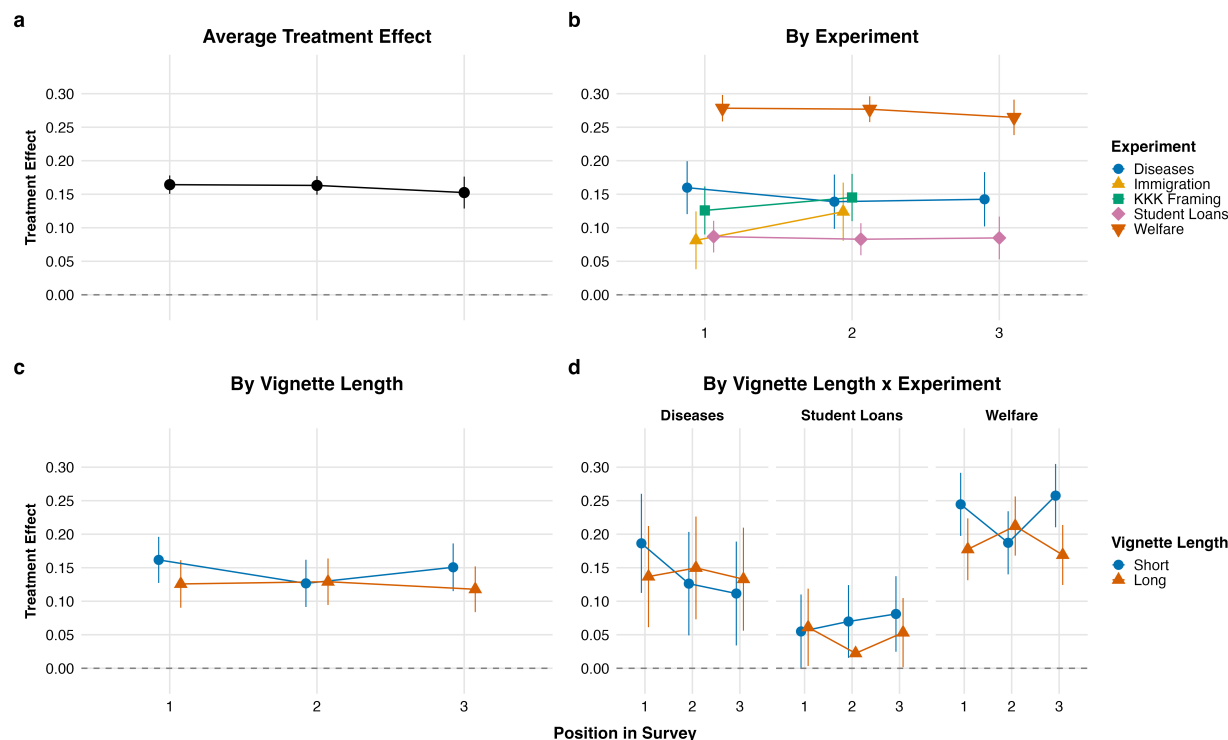


Figure 3: **Treatment effects by survey position.** a,b, Model-based adjusted treatment effects pooled across experiments (a) and by experiment (b); experiment-specific profiles are shown only at positions observed for that experiment. Estimates are predictive margins from a linear model with survey–experiment cell fixed effects and additive position and treatment-effect interactions by survey and experiment. c,d, Survey 5 treatment effects by vignette length, pooled across experiments (c) and by experiment (d), from a fully interacted treatment $\times$ position $\times$ length $\times$ experiment model. Outcomes were scaled 0–1 and oriented a priori so that positive values correspond to the treatment direction in the original study ($n = 5,030, 3,024$ and $3,678$ in Surveys 1, 3 and 5). Error bars show 95% confidence intervals based on respondent-clustered standard errors. Formal pairwise contrasts are reported in Supplementary Table 4, and an observed-support version of this figure in Supplementary Fig. 3.

In particular, treatment-effect contrasts for the welfare experiment, the study’s largest and most precisely estimated effect, were no larger than 0.2 percentage points in absolute value (all $P > 0.91$).

This conclusion did not depend on how contrasts were aggregated. Because Position 3 is observed in a different set of surveys than Positions 1 and 2, a pooled model with a single treatment-by-position interaction would blend within-survey position differences with between-survey differences in treatment effects; the stratified estimator avoids this by

construction, forming each contrast only within survey–experiment strata observed at both positions. Supplementary Table 7 compares these sample-size-weighted estimates with two alternatives: the additive-model contrast, which instead uses model structure to predict cells not observed by design, and an inverse-variance–weighted (IVW) variant that combines the same supported strata by inverse-variance weighting. The alternatives differed from the sample-size-weighted estimates by at most 1.2 and 0.5 percentage points, respectively, and under no approach did treatment effects vary systematically by position.

To distinguish small position effects from imprecisely estimated null effects, we computed ECIs for all treatment-effect contrasts (Supplementary Table 4). Pooled ECIs ranged from 1.7 to 3.2 percentage points; across the 11 experiment-specific contrasts, they ranged from 2.3 to 9.4 percentage points. All three pooled contrasts and 10 of the 11 experiment-specific contrasts met the equivalence criterion based on the conventional small-effect benchmark of $\pm 0.20\sigma_Y$ after BH adjustment across the full 14-contrast family. The sole exception was the immigration Position 2-versus-1 contrast, whose ECI was $0.24\sigma_Y$.

The design of Survey 5 additionally allowed us to examine whether position patterns differed between short and long versions of each experiment (Fig. 3c,d). If attentiveness declines over the course of a survey, longer vignettes might be more susceptible to position effects because they place greater demands on respondent engagement. Pooled across experiments, short-vignette effects showed a modest mid-survey dip before recovering (16.2 $\rightarrow$ 12.7 $\rightarrow$ 15.1 percentage points), whereas long-vignette effects remained relatively stable across positions (12.6 $\rightarrow$ 12.9 $\rightarrow$ 11.8). The direction of the small fluctuations varied across experiments and length conditions. For example, the welfare effect followed a U-shaped pattern for short vignettes (24.5 $\rightarrow$ 18.7 $\rightarrow$ 25.8) but an inverted-U pattern for long vignettes (17.8 $\rightarrow$ 21.2 $\rightarrow$ 16.9). A joint Wald test of all treatment-by-position-by-length interactions was not statistically significant ($\chi^2_6 = 8.89$, $P = 0.180$), providing little evidence that longer-vignette experiments were systematically more susceptible to position effects. Because these comparisons split Survey 5’s sample, however, they were estimated less precisely than the

main position contrasts.

Robustness to attrition and professionalism

A natural concern is that the overall stability of our estimates reflects the composition of our samples. We examined two such possibilities in turn: within-survey attrition and respondents' survey-taking experience (professionalism).

First, we considered attrition. Placing study items later in a survey exposes them to respondent dropout, which affects both sample size and, potentially, sample composition. Attrition reduces statistical power by shrinking the effective sample, and, if less attentive respondents are more likely to drop out, the sample remaining at later positions may be of higher quality. Because the preceding analyses were restricted to respondents who completed the full survey, we assessed whether including available responses from non-completers who had reached each item changed the magnitude or pattern of the estimates.

Across the five surveys, attrition rates ranged from 15.4% (Survey 1) to 28.2% (Survey 2), with a median of 24.1% and a pooled rate of 21.7%. Figure 4a plots the proportion of respondents remaining as a function of median survey time for each survey, with numbered markers indicating the positions at which data-quality measures were administered. Most attrition occurred within the first few minutes; by Position 2, most respondents who would eventually leave the survey had already done so.

To evaluate the sensitivity of our results to attrition, we re-estimated the adjusted position profiles using all available responses from respondents who passed initial screening, regardless of whether they completed the survey. Respondents, including non-completers, contributed data at a given position if they reached and answered the relevant item. Figure 4b shows the available-response minus completer-only estimate at each position; values near zero indicate that restricting the analysis to completers had little influence on the estimate. The two nested samples were fitted jointly, with each respondent keeping a single cluster identifier across samples, so that the standard errors of the between-sample differences

account for completers contributing to both samples.

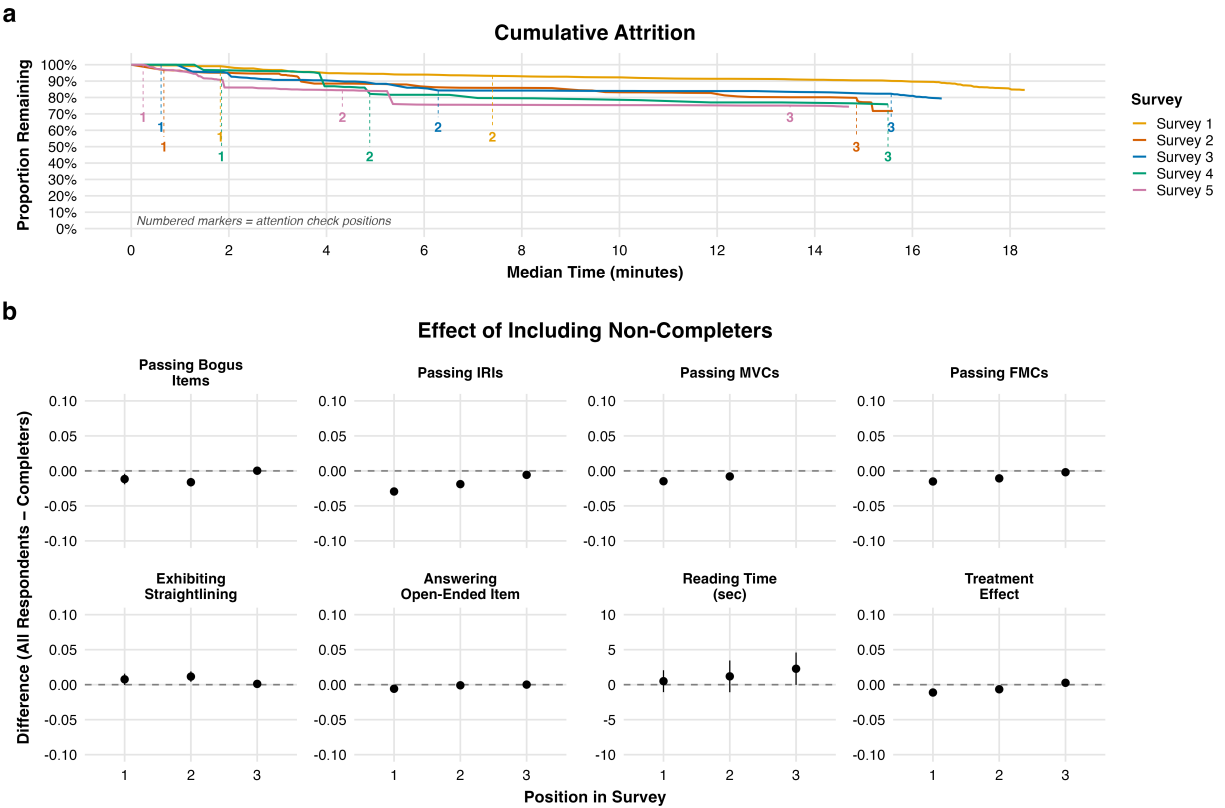


Figure 4: Respondent attrition and its effect on results. **a**, Proportion of respondents remaining as a function of median survey time; numbered dashed lines mark the positions of the data-quality measures. **b**, Difference between available-response and completer-only adjusted estimates at each position ($N = 19,370$ and $15,159$). The two nested samples are estimated jointly, with every coefficient of the underlying additive model allowed to differ by sample. Error bars show 95% confidence intervals based on respondent-clustered standard errors. Formal tests of whether including non-completers changes the pairwise position contrasts are reported in Supplementary Table 5, and an observed-support version of panel **b** in Supplementary Fig. 4.

Including non-completers shifted estimated levels most at the beginning of the survey. For example, adjusted IRI pass rates were 2.9 percentage points lower in the available-response sample at Position 1, compared with 1.9 points lower at Position 2 and 0.6 points lower at Position 3. Pooled treatment effects differed between samples by -1.1 , -0.7 and $+0.3$ percentage points at Positions 1–3. Formal comparisons asked whether the position contrast itself changed when non-completers were included. Several small differences were

statistically detectable. For example, the IRI Position-3-versus-1 contrast was 2.4 percentage points larger in the available-response sample (95% CI 1.9 to 2.8; BH-adjusted $P < 0.001$), and the pooled treatment-effect contrast was 1.3 percentage points larger (95% CI 0.5 to 2.1; BH-adjusted $P = 0.006$). Across all 16 attrition-sensitivity contrasts, however, ECIs were no larger than 2.8 percentage points for binary data-quality measures, 3.4 s for reading time and 2.0 percentage points for treatment effects; all met the $\pm 0.20\sigma_Y$ equivalence criterion after BH adjustment (Supplementary Table 5). Thus, including non-completers modestly reduced some early-position estimates but did not change the substantive interpretation of the position patterns, although this analysis does not reveal how respondents would have answered items they never reached.

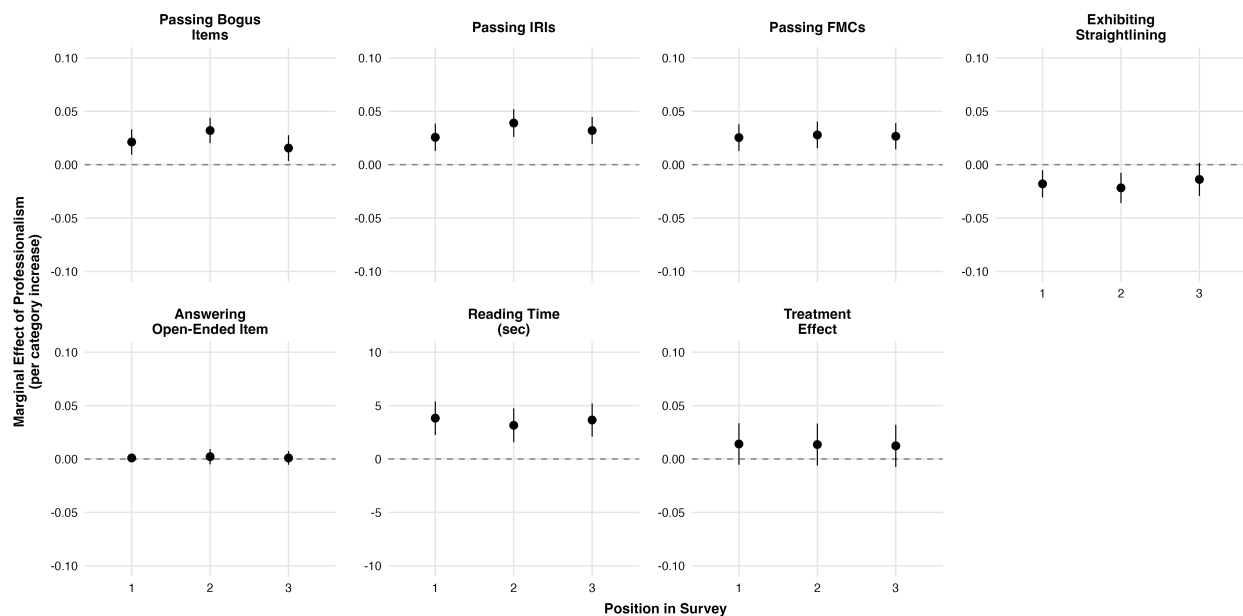


Figure 5: Survey-taking professionalism and data quality by survey position. Model-based average marginal effects of a one-category increase in the number of online surveys completed during the previous two weeks, for bogus-item, IRI and FMC pass rates, straightlining, substantive open-ended responses, vignette reading time and treatment effects (Surveys 4 and 5, $N = 1,414$ and $3,678$, as applicable). Data-quality models allow the professionalism slope to vary additively by position, survey and item; the treatment-effect panel uses a fully interacted treatment $\times$ professionalism $\times$ position $\times$ experiment model within Survey 5. Error bars show 95% confidence intervals based on respondent-clustered standard errors. Formal pairwise contrasts are reported in Supplementary Table 6.

Second, we considered professionalism. Respondents on opt-in platforms vary widely

in how many surveys they take, and our samples may include many experienced or professional survey-takers who are accustomed to long questionnaires and better able to sustain attention. A less experienced sample might therefore exhibit larger position effects. We measured professionalism as the number of online surveys completed in the previous two weeks (coded 0–4; see Methods) and estimated its association with each outcome at each position (Fig. 5). More professional respondents generally passed checks at higher rates, straightlined less often and spent longer reading vignettes. However, the professionalism slope did not vary systematically across positions. None of the 14 formal pairwise slope contrasts was statistically significant (all unadjusted $P \geq 0.11$; all BH-adjusted $P = 0.994$), and all met the $\pm 0.20\sigma_Y$ equivalence criterion, with ECIs no larger than $0.070\sigma_Y$ (Supplementary Table 6).

The same conclusion held when we relaxed the linearity assumption and treated professionalism as five categories. Joint tests of all position-by-category interactions were not statistically significant for any measure, before or after BH adjustment across the seven panels (all unadjusted $P \geq 0.10$; all adjusted $P \geq 0.52$). Together, the attrition and professionalism analyses indicate that the main findings are not driven by the exclusion of non-completers or by systematically different position effects among more- and less-professional respondents.

Discussion

Our results suggest that researchers need not treat item placement as a major threat to data quality or treatment-effect estimates, especially in surveys similar in length and design to those featured here. Across five surveys, seven data-quality measures and five embedded experiments, position differences in data quality were generally modest and treatment-effect estimates were stable. The data-quality differences we did detect did not point in a common direction: bogus items, FMCs, straightlining and open-ended responses indicated slightly lower data quality at later positions, whereas IRIs and MVCs indicated slightly higher data quality. Measures that place similar demands on respondents also did not move together:

FMCs and MVCs both require recall but showed opposite patterns, while straightlining and reading time, both related to response speed, likewise diverged.

The divergence across measures is difficult to reconcile with a general process of fatigue or warm-up. The Survey 1 MVC analysis instead points to a possible role for earlier module exposure: the later-position improvement was concentrated among respondents who had completed a different MVC module earlier, whereas performance at first exposure was nearly identical across positions. That comparison does not identify what aspect of the earlier module produced the difference, and prior exposure cannot account for every pattern. In Survey 4, where each respondent encountered only one check, the Color IRI improved across positions and the Web IRI improved at Position 2. Together, these findings suggest that the observed position effects varied with the measure and survey context rather than reflecting a uniform change in respondent engagement. Distinguishing among these possibilities will require designs that independently manipulate prior check exposure, check format and perceived monitoring—for example, by holding the amount of preceding survey material constant while varying whether respondents encounter an earlier check of the same or a different format and whether monitoring is made explicit.

Our finding that position effects were generally modest and directionally inconsistent aligns with the mixed picture in the existing literature, where evidence for systematic declines in attentiveness has been limited and sensitive to the measures employed^{6;9;10}. Differences in measurement, item format and survey design may help explain why prior studies have reached divergent conclusions, consistent with broader concerns about generalizing from narrow sets of stimuli or measures²⁷. Selective publication of statistically significant findings may also contribute to the apparent inconsistency^{28;29}. Our study substantially broadens this evidence base by evaluating seven data-quality measures, several represented by multiple items and formats, across five complementary randomized designs. It also provides, to our knowledge, the first systematic multi-experiment test of whether placing an experiment early, midway through, or late in a survey changes its estimated treatment effect.

All five surveys were fielded on Lucid using US samples, and the results may not generalize to other platforms or populations. Yet, because Lucid samples have exhibited lower average response quality than samples from several other online platforms¹⁵, we believe our design represents a favourable scenario for detecting position effects. Thus, the fact that we find little systematic variation suggests that such effects are unlikely to be substantially greater on higher-quality platforms.

It is also worth noting that Surveys 2–5 required respondents to pass a CAPTCHA before beginning the survey, and Surveys 2–4 included a pre-survey LLM-detection check. These procedures may have removed some of the least attentive or automated respondents at the outset and thereby attenuated subsequent position differences. At the same time, such screening reflects common quality-control practice, making our estimates directly relevant to surveys using comparable procedures.

Additionally, our design compares three broad survey locations rather than tracing continuous changes in attention, and the attrition analysis cannot identify how respondents who dropped out would have answered items they never reached. Finally, we manipulated where a measure appears within the questionnaire, not the content of the items immediately preceding it; our results therefore concern position in a survey, not the question-order or carryover effects that arise when specific items precede one another.

For researchers fielding experiments within omnibus surveys, our findings carry a practical message that within surveys of similar length and design, where an experiment is placed has little bearing on the estimates it yields. Experiments placed towards the end of a survey generally yielded substantively similar treatment-effect estimates to those placed near the beginning. These findings should be particularly reassuring for resource-constrained researchers who share survey space with other modules or research teams, where placement is often dictated by logistical rather than methodological considerations. Best practices for maintaining data quality, including pre-survey screening and restricting surveys to a reasonable length, remain important, of course. But researchers can be confident that, in

surveys like those studied here, survey position is unlikely to substantially or systematically affect data quality or estimated treatment effects.

Methods

Ethics and reproducibility

The study protocols governing all data collections were reviewed by the institutional review boards at New York University (IRB-FY2020-4097) and the Massachusetts Institute of Technology (E-6543 and E-6289) and determined to be exempt from the requirement for formal ethics review. The research complied with all relevant ethical obligations. All participants provided informed consent before taking part and were compensated at the platform-standard rate. Respondents were not deceived about the purpose of the study; experimental vignettes described hypothetical scenarios adapted from published studies, and respondents were debriefed. An analysis plan was registered on the Open Science Framework (OSF; <https://osf.io/b5wyr/>) after Surveys 1–4 had been fielded and before data collection for Survey 5 was complete; the subsection ‘Preregistration and deviations’ below describes the timing, scope and deviations in detail.

Study overview

We fielded five independent surveys on the Lucid platform between February 2020 and August 2025 (total $N = 15,159$ completed responses). Each survey randomly assigned a range of data-quality measures, including instructed-response items, comprehension checks and behavioural indicators, to different positions within the questionnaire to test whether data quality varies as respondents progress through a survey. Three surveys additionally embedded previously published survey experiments at randomized positions to test whether treatment-effect estimates vary by placement. The designs ranged from fully within-subjects to pure between-subjects, allowing us to distinguish position effects from within-survey learn-

ing; the Design subsection below describes each survey in detail. Unless otherwise noted, the analyses are restricted to respondents who completed the full survey; sensitivity to this completer-only restriction is examined separately in ‘Robustness to attrition and professionalism’ in the Results.

Samples

All surveys were fielded on Lucid Theorem, which recruits opt-in samples using quotas for age, gender, race/ethnicity and geographic region to approximate US Census demographics¹⁸. Lucid and similar non-probability platforms have become common sources of participants for survey-based research, particularly among researchers facing resource constraints³⁰, making our results relevant to conditions under which much of this research is conducted. Detailed demographic characteristics of each sample are provided in Supplementary Table 1. Median completion times were similar across the five surveys (14.7–18.3 min), although mean durations varied more widely (22.3–38.9 min) owing to right-skewed distributions driven by respondents who left the survey idle before returning to complete it (Table 1 and Supplementary Table 2).

Before encountering the study materials, respondents completed screening procedures that varied across surveys. Survey 1 required informed consent but no additional screening. Surveys 2–5, fielded later amid growing concerns about automated and LLM-generated responses, added a CAPTCHA and an open-ended screening item. The latter presented a visible question alongside a hidden instruction, rendered in white text, directing the reader to enter a specific numeric string; responses following the hidden instruction were flagged as likely automated, whereas respondents answering the visible question were permitted to continue. Across the five surveys, 19,370 respondents consented and, in Surveys 2–5, answered the CAPTCHA correctly; they form the available-response sample used in the attrition analysis. Of these, 15,159 respondents (78.3%) completed the survey and form the main analytic sample. In Survey 5, the LLM-detection item was moved from the pre-survey screen to one of

three randomized positions within the questionnaire, where it was analysed as a data-quality measure.

Data-quality measures

We use *data-quality measures* as the umbrella term for the seven outcome types examined in this study. Four were performance-based checks intended to capture attentiveness or related aspects of survey engagement. *Instructed-response items* (IRIs) embed a directive within a prompt and require respondents to override their natural response; for example, “Please select ‘slightly agree’ for this item”¹. IRIs include both *screeners*, which present the directive as a relatively short, standalone instruction, and *instructional manipulation checks* (IMCs), which embed the directive within a longer block of text³¹. *Factual manipulation checks* (FMCs) assess whether respondents correctly recall information presented in experimental materials¹⁹, and *mock vignette checks* (MVCs) assess comprehension of practice vignettes before treatment assignment¹⁷. *Bogus items* present implausible or verifiable statements, such as “I have never brushed my teeth”, to identify careless responding.

These performance-based checks do not necessarily measure a single construct. IRIs primarily capture whether respondents follow explicit instructions, especially when those instructions are counterintuitive or request untrue responses. FMCs and MVCs additionally depend on respondents’ ability to comprehend and recall information from an earlier screen, whereas bogus items capture, at least in part, whether respondents provide plausible and earnest responses.

The remaining three measures were behavioural indicators: *reading time*, measured as time spent reading treatment vignettes; *straightlining*, defined as identical responses across a battery of grid items; and the *substantive open-ended response*. These measures do not require respondents to follow or recall an explicit instruction. Instead, they capture aspects of how respondents engage with survey content, including time spent reading, differentiation among closed-ended responses and willingness to provide a substantive written answer.

Overall, the data-quality measures differ in format, difficulty and the behaviour they capture. This diversity allows us to assess whether position effects generalize across distinct components of data quality. Within each measure type, we also included multiple item variants, allowing us to test whether position effects generalize across items of the same type or depend on item-specific features such as difficulty and susceptibility to learning. Table 2 describes each measure and indicates the surveys in which it was used; exact item wordings are provided in Supplementary Note 2.

Design

Across the five surveys, data-quality measures were administered at different positions within the questionnaire: near the beginning (Position 1; generally after consent), approximately halfway through (Position 2) or near the end (Position 3; before the debrief). Surveys 1 and 2 compared two positions, whereas Surveys 3–5 used all three, enabling direct tests of non-monotonic position patterns. The surveys used complementary designs to distinguish position effects from repeated-exposure effects and to examine their generalizability across measures and experimental formats (Table 1).

Survey 1 ($n = 5,030$) used a two-position within-subjects design in which each respondent completed two experimental rounds at Positions 1 and 2; the second half of the survey consisted of unrelated modules. Each round included an experiment followed by an FMC. When administered, a mock-vignette module comprising one vignette and three MVCs preceded the experiment. At Position 1, 20.5% of respondents were randomly assigned not to receive the module; the remaining respondents received it. All respondents were assigned to receive a mock-vignette module at Position 2. Respondents assigned modules at both positions encountered different mock vignettes. No substitute module was administered when the Position 1 module was omitted, so those respondents completed less survey material before reaching Position 2.

Survey 2 ($n = 2,013$) administered IRIs at the start (Position 1) and end (Position 3)

of the survey. The Color IRI was administered at both positions, whereas the Simple and Web IRIs were independently assigned to the beginning or end. Consequently, one to three IRI modules appeared at each position and were presented in randomized order.

Survey 3 ($n = 3,024$) randomized three IRI types and three experiments across three positions. Each position contained one IRI followed by one experiment; the identities of the IRI and experiment were randomized across positions, but their within-position order was fixed. Survey 4 ($n = 1,414$) used a pure between-subjects design in which each respondent encountered only one IRI, randomly assigned to one of the three positions, thereby isolating position effects from learning induced by repeated exposure.

Survey 5 ($n = 3,678$) crossed three positions with two experimental-vignette lengths (short versus long) to test whether vignette length moderated position effects. The Section IRI, Issue IRI and Brush bogus-item modules were randomly assigned one apiece across the three positions. Each position also contained one experiment and zero, one or two additional modules: a seven-item opinion grid, which contained the two Grid bogus items and was also used to measure straightlining, and an open-ended-response module. Within each position, the assigned IRI or bogus-item module and the experiment were presented in randomized order, with any additional module or modules treated as a separate block. When both additional modules appeared at the same position, their order within the block was also randomized.

All random assignment and order randomization were implemented using Qualtrics' built-in randomizer, configured to balance assignment across conditions.

Experiments

To test whether treatment-effect estimates vary by placement, we selected published survey experiments for which prior research had documented treatment effects across diverse policy domains and vignette structures (full materials are provided in Supplementary Note 2). The *welfare* experiment²⁰ asked respondents to consider a man on social welfare who was

described either as having been injured on the job (“unlucky”) or as fit and healthy but unmotivated (“lazy”), and then measured support for tightening eligibility requirements. The *disease* experiment²¹ presented the classic disease-risk problem, framing equivalent outcomes in terms of lives saved or lives lost. The *student loan* experiment²² presented information about rising student debt, with a treatment condition adding expert warnings about the negative economic consequences of a forgiveness program, and measured support for loan forgiveness. Survey 1 additionally included a *KKK rally* experiment²³, which framed a planned Ku Klux Klan rally on a university campus in terms of either free speech or public safety, and an *immigration* experiment²⁴, which varied the skill level and national origin of an immigrant applying for a US work permit and measured support for admitting the applicant.

For comparability across experiments, all outcomes were rescaled to a common 0–1 metric and oriented so that positive values corresponded to the treatment direction defined in the original study. This enables treatment-effect estimates to be pooled across experiments and compared on a common scale.

Statistical analysis

We distinguish the *additive models* used to plot model-based adjusted means (predictive margins) in the figures from the *two-stage stratified estimator* used for all formal pairwise position comparisons. All outcomes were analysed with linear models (linear probability models for binary outcomes); unless otherwise stated, covariance matrices were clustered at the respondent level to account for respondents contributing multiple observations.

For each pair of survey positions, the estimand for the data-quality measures is the average difference in outcome levels between the two positions, defined over the survey–item combinations in which both positions were fielded; for the experiments, it is the corresponding difference in average treatment effects. The surveys contributing data differ across positions (for example, Survey 1 fielded experiments at Positions 1 and 2 but not Position 3),

so estimating a single position effect, or a single treatment-by-position interaction, in the pooled data would conflate within-survey position differences with differences between surveys in outcome levels or treatment effects. We therefore estimated every contrast within survey, using only strata in which both positions were observed, and combined estimates across surveys only afterwards.

For Figs. 1 and 2, the additive model included a fixed effect for every observed survey-item combination, position indicators, position-by-survey interactions and position-by-item interactions. Survey and item main effects are absorbed by the survey-item cell fixed effects. We omitted the position-by-survey-by-item interaction, allowing position patterns to vary additively by survey and item while supplying model-based predictions for survey-item-position cells not observed by design, an additivity assumption that the formal comparisons below avoid. These plotted profiles are model-based adjusted means (predictive margins): we average the model's predictions over a fixed empirical survey-item composition, evaluated at each position in turn. Item- and survey-specific profiles were computed with the same model specification fitted to the relevant subset of the data and are displayed only at positions observed for that item or survey-item combination. In panels containing a single survey or item, the model is saturated, so the adjusted values equal observed cell means.

Separately, for Survey 1 MVCs, we decomposed the aggregate position difference into a comparison of first MVC-module exposures across Positions 1 and 2 and a comparison at Position 2 by prior MVC-module exposure. We fitted a linear probability model with indicators for Position 2 and prior MVC-module exposure, MVC-item fixed effects and respondent-clustered standard errors. The Position 2 coefficient compares first MVC-module exposures at Positions 2 and 1, whereas the prior-exposure coefficient compares Position 2 outcomes among respondents assigned and not assigned to an MVC module at Position 1.

For Fig. 3a,b, the analogous additive model included survey-experiment cell fixed effects; position-by-survey and position-by-experiment interactions; treatment-by-survey and treatment-by-experiment interactions; and treatment-by-position interactions that could

577 vary additively by survey and experiment. Equivalently, the model included treatment-
578 by-position-by-survey and treatment-by-position-by-experiment terms but omitted the four-
579 way treatment-by-position-by-survey-by-experiment interaction. At each position, we com-
580 puted the adjusted treatment effect by predicting every reference observation under treat-
581 ment and control and averaging the individual prediction differences over a fixed survey-
582 experiment composition. Experiment-specific profiles show only positions observed for that
583 experiment. The Survey 5 vignette-length analysis used a fully interacted treatment $\times$ po-
584 sition $\times$ length $\times$ experiment model because all required cells were observed. We tested
585 whether vignette length moderated the treatment-by-position pattern with a respondent-
586 clustered joint Wald test of all treatment-by-position-by-length interaction terms, including
587 their experiment-specific components.

588 Formal position comparisons used a two-stage stratified estimator. Within each sur-
589 vey, we fitted fully saturated models over the observed cells. Data-quality models estimated
590 a separate mean for every item-position cell, and treatment models estimated a separate
591 intercept and treatment effect for every experiment-position cell. For each position pair,
592 we retained only survey-item or survey-experiment strata in which both positions were
593 observed and formed the later-minus-earlier contrast within each supported stratum. We
594 then combined these stratum-specific contrasts using sample-size weights. Each stratum was
595 weighted by its number of observations in the pooled analytic sample, with weights renor-
596 malized to sum to one over the strata supporting each comparison. Because each pooled
597 contrast is a linear combination of cell estimates, its standard error was computed from the
598 full respondent-clustered covariance matrix of those estimates within each survey. Estimates
599 from different surveys, which were fielded independently, were treated as uncorrelated, so
600 survey-specific variance contributions were summed. Consequently, each pairwise contrast is
601 defined over its own set of supported strata and need not equal the simple difference between
602 the plotted adjusted means. Because the supported strata and their normalized weights can
603 differ across position pairs, these pooled pairwise comparisons are not constrained to be

transitive. For example, the Position-3-versus-1 estimate need not equal the sum of the Position-3-versus-2 and Position-2-versus-1 estimates. In the main analyses, conventional P values, BH adjustments and equivalence tests all use these stratified comparisons.

For comparison, Supplementary Table 7 reports two alternatives. The additive-model contrast is derived from the additive model used for the figures, relying on its additive structure to predict cells not observed by design. The inverse-variance-weighted (IVW) contrast uses the same supported strata but combines survey-specific estimates by fixed-effects (inverse-variance) meta-analysis rather than by sample-size weighting. The IVW and sample-size-weighted contrasts are both computed separately for each position pair; because their supported populations and weights can change by pair, neither is required to satisfy transitivity. These comparisons assess sensitivity to extrapolation and cross-survey weighting.

All reported 95% confidence intervals are Wald intervals. Unless explicitly labelled as BH-adjusted, P values in the main text are unadjusted. We applied BH adjustment separately within four families: 13 data-quality contrasts, 14 treatment-effect contrasts, 16 attrition-sensitivity contrasts and 14 professionalism contrasts.

To assess whether position differences were bounded within a substantively small range rather than merely failing to differ significantly from zero, we conducted two one-sided equivalence tests (TOST), in which the null hypothesis is a difference of at least a given margin and rejection provides positive evidence that any difference is smaller²⁶. We used the margin $\pm 0.20\sigma_Y$, where σ_Y denotes the standard deviation of the relevant outcome; this is the conventional small-effect benchmark²⁵ and is more stringent than the $\pm 0.36\sigma_Y$ default suggested in the equivalence-testing literature²⁶. TOST P values were BH-adjusted within the same four families as the conventional P values; because rejection, not failure to reject, constitutes the evidence of equivalence, such corrections can be applied directly to equivalence-test P values and are conservative for equivalence claims²⁶. For each contrast we also computed the equivalence confidence interval (ECI), reported by its margin

δ : the smallest value such that all effects larger than δ in magnitude can be statistically ruled out at $\alpha = 0.05$, equal to the larger absolute endpoint of the 90% confidence interval (Supplementary Note 3).

For the attrition analysis (Fig. 4), we created stacked copies of the available-response and completer-only data and included an indicator for sample membership interacted with every term in the corresponding additive model. This allowed every coefficient, including baseline cell means and position patterns, to differ between the two samples. It is algebraically equivalent to fitting separate additive models to the two samples, but joint estimation preserves their covariance because completers contribute to both samples. Respondents retained the same cluster identifier in both copies. Formal attrition-sensitivity inference instead used paired, fully saturated observed-cell models. Within each sample we formed the stratified position contrast on strata observed at both positions and in both samples, then tested the available-response minus completer difference in that contrast while retaining their covariance.

For the professionalism analysis (Fig. 5), professionalism was measured using five response categories for the number of online surveys completed during the previous two weeks, from ‘none’ to ‘more than 20’, coded 0–4. The data-quality additive models added professionalism and interactions allowing its slope to vary additively by position, survey and item, including professionalism-by-position-by-survey and professionalism-by-position-by-item terms. The treatment-effect panel, observed only in Survey 5, used a fully interacted treatment $\times$ professionalism $\times$ position $\times$ experiment model. Formal pairwise stratified comparisons estimated a separate professionalism slope in every observed item–position cell (and a separate treatment-by-professionalism slope in every experiment–position cell), then contrasted slopes across positions using sample-size weights. To relax the linearity assumption, we re-estimated the models treating professionalism as a five-category factor and jointly tested all position-by-category interactions using respondent-clustered Wald tests.

All analyses were conducted in R version 4.5.2. The replication archive provides the

analysis code, package versions and materials needed to reproduce the reported results.

Preregistration and deviations

The registration specified eight hypotheses (H1–H5, H6a, H6b and H7). H1 predicted that attention-check performance would vary by survey position, while H3–H5 made analogous predictions for factual manipulation checks, vignette reading time and open-ended response quality; the latter was registered in terms of response length and elaboration and is operationalized here as the binary substantive-response measure described in Table 2. For presentation in the manuscript, we group H1 and H3–H5 under the broader hypothesis that data quality depends on survey position (manuscript H1), while registered H2, which predicted that treatment effects would vary by experiment position, corresponds to manuscript H2. H6a and H6b, the only directional hypotheses, predicted better performance on the second presentation of the same item (within-item learning) and on later, distinct checks following earlier check exposure (cross-item learning), respectively. H6a is addressed by the repeated Color item in Survey 2, whereas H6b is addressed by the disaggregated IRI analysis (Fig. 2). The Survey 1 MVC decomposition provides an additional examination of H6b within a single measure type, comparing Position 2 performance among respondents with and without prior exposure to a different MVC module. H7, that complex and simple attention checks would show different position effects, is addressed by the item-level heterogeneity in Fig. 2b.

The hypotheses and analysis plan were registered on the Open Science Framework while data collection for Survey 5 was ongoing, after Surveys 1–4 had been completed and exploratory analyses of those surveys had been conducted. As disclosed in the registration, we had access to the partially collected Survey 5 data but had not examined them and did not do so until data collection was complete. The registration should therefore be understood as retrospectively documenting the hypotheses and analyses applied to Surveys 1–4 while prospectively specifying the analysis of Survey 5 with respect to inspection of its outcomes.

684 Because the primary registered hypotheses concerning position and treatment-effect hetero-
685 geneity were non-directional and the study's conclusions rest on estimated magnitudes and
686 equivalence bounds rather than on confirming a predicted pattern, exposure to Surveys 1–4
687 before registration could not have meaningfully shaped them.

688 The registration used different survey labels and listed approximate sample sizes; the
689 correspondence between those labels and the survey numbering used here, together with
690 registered and final analytic sample sizes, is reported in Supplementary Table 1. The regis-
691 tration incorrectly listed Survey 1 as containing approximately 2,000 respondents; the final
692 completed sample was 5,030. Survey 5's target of 4,000 yielded 3,678 completed responses
693 after the completion and screening criteria were applied. Four experimental modules men-
694 tioned under Survey 2 belonged to unrelated projects sharing the omnibus survey and are
695 not analysed here.

696 Although the registration did not include a formal sample-size rationale, the design
697 of Survey 5 was informed by a priori power calculations conducted before its data collection,
698 using treatment-effect estimates and outcome variances from Survey 1 and published repli-
699 cation data to determine the sample required to detect each experiment's effect with 80%
700 power at $\alpha = 0.05$, allowing for attenuated effects and for the share of respondents encoun-
701 tering each experiment at a given position. These calculations also guided the selection of
702 experiments (power-analysis code is included in the replication archive).

703 The registration specified descriptive statistics by position, pairwise comparisons with
704 adjustment for multiple testing, meta-analytic combination of effects across surveys and ex-
705 clusion of respondents who did not complete the survey or provide the relevant outcome.
706 The analyses reported here follow each element of this plan. The registration left the combi-
707 nation weights unspecified, however, and the choice matters here because different position
708 pairs are supported by different survey–item or survey–experiment strata. We combine sup-
709 ported strata using the sample-size weights described above, whereas a conventional reading
710 of a meta-analytic combination implies inverse-variance weighting. Supplementary Table 7

therefore reports the IVW estimates alongside ours; the two yield the same substantive conclusions. The registration likewise did not specify the exact regression parameterization, the restriction to strata observed at both positions or the multiple-testing procedure; these, like the weights, are implementation choices within the registered plan rather than departures from it. Supplementary Table 7 assesses the consequences of the parameterization and the supported-strata restriction, and P values are reported both with and without BH adjustment.

The equivalence tests and ECIs, the analysis including available responses from non-completers, the position-by-professionalism analyses, the categorical specification of professionalism and the comparison of alternative aggregation procedures were not specified in the registration and should be interpreted as supplementary or robustness analyses. Survey experience, which we use to operationalize professionalism, was included among the registered measured variables, but no hypothesis concerning its interaction with survey position was registered. Straightlining was also listed as a measured variable but was not associated with a separately registered hypothesis.

Data availability

Anonymized data for this project are available via the Open Science Framework at <https://osf.io/86jrk/>.

Code availability

Materials and analysis code for this project are available via the Open Science Framework at <https://osf.io/86jrk/>.

References

- [1] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58(3):739–753, 2014.
- [2] Jason L Huang, Mengqiao Liu, and Nathan A Bowling. Insufficient effort responding: examining an insidious confound in survey data. *Journal of applied psychology*, 100(3):828, 2015.
- [3] John V. Kane. More than meets the eye: a guide for anticipating and investigating nonsignificant results in survey experiments. *Journal of Experimental Political Science*, 12(1):110–125, 2025.
- [4] Scott Blatte and Brian Schaffner. Surviving the screens: the problem of hidden inattentive respondents in online surveys. *Political Science Research and Methods*, pages 1–16, 2026.
- [5] Stephen Ansolabehere and Douglas Rivers. Cooperative survey research. *Annual Review of Political Science*, 16(1):307–329, 2013.
- [6] Anthony M Gibson and Nathan A Bowling. The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, 36(2):410, 2020.
- [7] Nathan A. Bowling, Anthony M. Gibson, Joseph W. Houpt, and Cheyna K. Brower. Will the questions ever end? Person-Level increases in careless responding during questionnaire completion. *Organizational Research Methods*, 24(4):718–738, 2021. doi: 10.1177/1094428120947794. URL <https://doi.org/10.1177/1094428120947794>.
- [8] Mirta Galesic and Michael Bosnjak. Effects of questionnaire length on participation

and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2): 349–360, 2009.

[9] Martin Bäckström and Fredrik Björklund. Is reliability compromised towards the end of long personality inventories? *European Journal of Psychological Assessment*, 35(1): 14, 2019.

[10] Kirk Bansak, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. The number of choice tasks and survey satisficing in conjoint experiments. *Political Analysis*, 26(1): 112–119, 2018.

[11] Tobias Gummer, Joss Roßmann, and Henning Silber. Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, 50(1):238–264, 2021.

[12] Franki YH Kung, Navio Kwok, and Douglas J. Brown. Are attention check questions a threat to scale validity? *Applied Psychology*, 67(2):264–283, 2018.

[13] Moreno Mancosu, Riccardo Ladini, and Cristiano Vezzoni. ‘short is better’. evaluating the attentiveness of online respondents through screener questions in a real survey environment. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 141(1):30–45, 2019.

[14] Adam J. Berinsky, Alejandro Frydman, Michele F. Margolis, Michael W. Sances, and Diana Camilla Valerio. Measuring attentiveness in self-administered surveys. *Public Opinion Quarterly*, 88(1):214–241, 2024.

[15] Michael N. Stagnaro, James N. Druckman, Adam J. Berinsky, Antonio A. Arechar, Robb Willer, and David G. Rand. Representativeness and response validity across nine opt-in online samples. *Nature Human Behaviour*, pages 1–14, 2026.

- [16] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. Can we turn shirkers into workers? *Journal of Experimental Social Psychology*, 66:20–28, 2016. ISSN 0022-1031. doi: <https://doi.org/10.1016/j.jesp.2015.09.010>. URL <https://www.sciencedirect.com/science/article/pii/S0022103115001201>. Rigorous and Replicable Methods in Social Psychology.
- [17] John V. Kane, Yamil R. Velez, and Jason Barabas. Analyze the attentive and bypass bias: mock vignette checks in survey experiments. *Political Science Research and Methods*, 11(2):293–310, 2023.
- [18] Alexander Coppock and Oliver A. McClellan. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, 6(1):2053168018822174, 2019.
- [19] John V. Kane and Jason Barabas. No harm in checking: Using factual manipulation checks to assess attentiveness in experiments. *American Journal of Political Science*, 63(1):234–249, 2019.
- [20] Lene Aarøe and Michael Bang Petersen. Crowding out culture: Scandinavians and Americans agree on social welfare in the face of deservingness cues. *Journal of Politics*, 76(3):684–697, 2014.
- [21] Daniel Kahneman and Amos Tversky. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.
- [22] Kevin J. Mullinix, Thomas J. Leeper, James N. Druckman, and Jeremy Freese. The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2):109–138, 2015.
- [23] Thomas E. Nelson, Rosalee A. Clawson, and Zoe M. Oxley. Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, 91(3):567–583, 1997.

- [24] Nicholas A. Valentino, Stuart N. Soroka, Shanto Iyengar, Toril Aalberg, Raymond Duch, Marta Fraile, Kyu S. Hahn, Kasper M. Hansen, Allison Harell, Marc Helbling, Simon D. Jackman, and Tetsuro Kobayashi. Economic and cultural drivers of immigrant support worldwide. *British Journal of Political Science*, 49(4):1201–1226, 2019.
- [25] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.
- [26] Erin Hartman and F. Daniel Hidalgo. An equivalence approach to balance and placebo tests. *American Journal of Political Science*, 62(4):1000–1013, 2018.
- [27] Tal Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2022.
- [28] Philip Moniz, James N. Druckman, and Jeremy Freese. The file drawer problem in social science survey experiments. *Proceedings of the National Academy of Sciences*, 122(12):e2426937122, 2025.
- [29] Tamkinat Rauf, Jan Gerrit Voelkel, James Druckman, and Jeremy Freese. An audit of social science survey experiments. *Public Opinion Quarterly*, 89(4):1063–1086, 2025.
- [30] Jennifer Jerit and Jason Barabas. Are nonprobability surveys fit for purpose? *Public Opinion Quarterly*, 87(3):816–840, 09 2023. ISSN 0033-362X. doi: 10.1093/poq/nfad037. URL <https://doi.org/10.1093/poq/nfad037>.
- [31] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, 2009.

Acknowledgements

We thank our discussant L. Jenke and participants at the 2026 Midwest Political Science Association annual meeting for helpful comments, and E. Parker for excellent research assistance. We also thank G. Péloquin-Skulski, J. Loffredo, and A. Frydman for their assistance

826 with fielding the Political Experiments Research Lab omnibus surveys. Funding for this
827 project was provided by the Department of Political Science at the Massachusetts Institute
828 of Technology (A.J.B.). The funder had no role in study design, data collection and analysis,
829 decision to publish or preparation of the manuscript.

830 **Author contributions**

831 J.J., A.J.B. and J.V.K. designed the research; A.J.B. secured funding; J.J. and J.V.K. col-
832 lected the data; J.J. led the project and analysed the data; J.J. and J.V.K. wrote the paper;
833 N.S. reviewed the literature and conducted preliminary analyses for the professionalism sec-
834 tion; all authors revised and approved the final manuscript.

835 **Competing interests**

836 The authors declare no competing interests.