

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 where researchers place items and experiments within a survey is unlikely to matter substantially for data quality or for the treatment effects they estimate.

20 Surveys are among the most widely used instruments for studying human attitudes
21 and behaviour. Yet, the validity of many survey measures depends on respondents' effort-
22 fully reading and considering the questions presented to them. When respondents engage
23 carelessly with those questions, the consequences can be substantial. Careless responding
24 can distort estimates of population means and correlations between items, and attenuate
25 estimated treatment effects in survey experiments.¹⁻³

26 Researchers commonly use attention checks to “screen out” low-effort respondents
27 before the survey begins, but passing such a screener does not ensure that a respondent
28 remains sufficiently attentive for the duration of the study.⁴ Potential variation in attentive-
29 ness raises concerns for all survey research, but the concern may be especially pronounced
30 in multi-investigator omnibus surveys, which are often relatively long and contain several
31 unrelated modules or experiments. Researchers facing resource constraints may include sev-
32 eral experiments within a single study, and separate research teams routinely submit items
33 to one large shared survey, as in cooperative multi-investigator studies.⁵ Although these
34 practices use resources efficiently, the longer surveys they produce raise important practical
35 questions. To what extent does data quality—as measured by, for example, attention to
36 particular items, response times or willingness to answer open-ended questions—change as
37 a survey progresses? Are survey components such as individual questions, question grids
38 or vignette-based experiments systematically disadvantaged when placed at the beginning,
39 middle, or end of a survey?

40 Despite the widespread reliance on lengthy, omnibus surveys, researchers currently
41 lack consistent, robust evidence on these critical questions. On one hand, several stud-
42 ies report significant variability in data quality over the course of a survey. Gibson and
43 Bowling,⁶ for example, find some support for the hypothesis that survey length affects the
44 likelihood of providing correct and consistent responses, as do Bowling et al.,⁷ although
45 the latter results depend somewhat on the particular measure of data quality employed.
46 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 samples,⁶ 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 assessed.¹⁴ 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 survey,¹⁰ nor did they use enough survey positions to detect *non-monotonic* variation in data quality.⁹ 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 carelessly.⁸

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 vary.¹ 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 measures.¹⁶ This matters because inattentiveness will typically attenuate treatment effect estimates.¹⁷ 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 demographics.¹⁸ Because Lucid samples have exhibited lower average response quality than samples from several other online platforms,¹⁵ 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 follow;¹ bogus items, which present implausible or verifiable statements; factual manipulation checks (FMCs), which test recall of experimental content;¹⁹ mock vignette checks (MVCs), which test comprehension of practice vignettes;¹⁷ and three behavioral 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, 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: ^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. Experiment sources: “Student loan”;²⁰ “KKK”;²¹ “Welfare”;²² “Immigration”;²³ “Disease”.²⁴

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 experiment,²² the classic disease-risk framing problem,²⁴ a student-loan framing experiment,²⁰ a KKK rally experiment²¹ and an immigration experiment.²³ 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 benchmark,²⁵ 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 better 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 substantial 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.

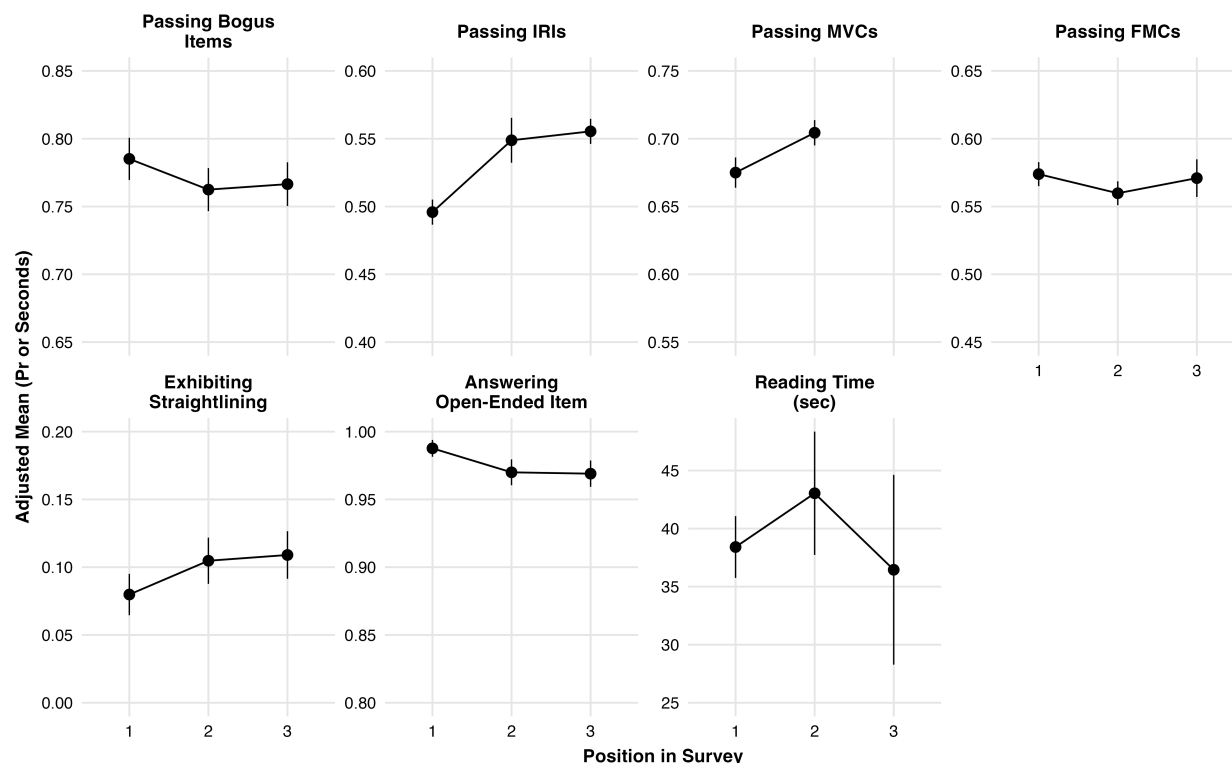


Figure 1: **Data quality by survey position.** Model-based adjusted means (predictive margins) for bogus-item, IRI, MVC and FMC pass rates, straightlining, substantive open-ended responses and vignette reading time (total $n = 15,159$). Estimates are from linear models with survey-item cell fixed effects 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, similarly improved from Position 1 to 2 (+2.9 percentage points, 95% CI 1.9 to 3.9, $P < 0.001$). Given the format of Survey 1, however, this improvement could reflect learning that factual

questions followed the vignettes, leading respondents to attend more closely to the second vignette. Corroborating this point, the contrast between the FMC and MVC patterns is notable because both measures require attentiveness, comprehension and recall, suggesting that the MVC improvement does not reflect a general increase in data quality.

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 tests,²⁶ 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),

182 a symmetric interval $\pm\delta$ that we report by its margin δ : the smallest value such that all
 183 effects larger than δ in magnitude can be statistically ruled out at $\alpha = 0.05$. Smaller ECIs
 184 therefore indicate tighter bounds on possible position effects. Across the 13 contrasts, ECIs
 185 ranged from 2.4 to 7.0 percentage points for binary data-quality measures and from 2.0 to
 186 9.5 s for vignette reading time. Using the conventional small-effect benchmark of $\pm 0.20\sigma_Y$
 187 (ref. 25), all 13 contrasts were statistically equivalent after BH adjustment (Supplementary
 188 Table 3).

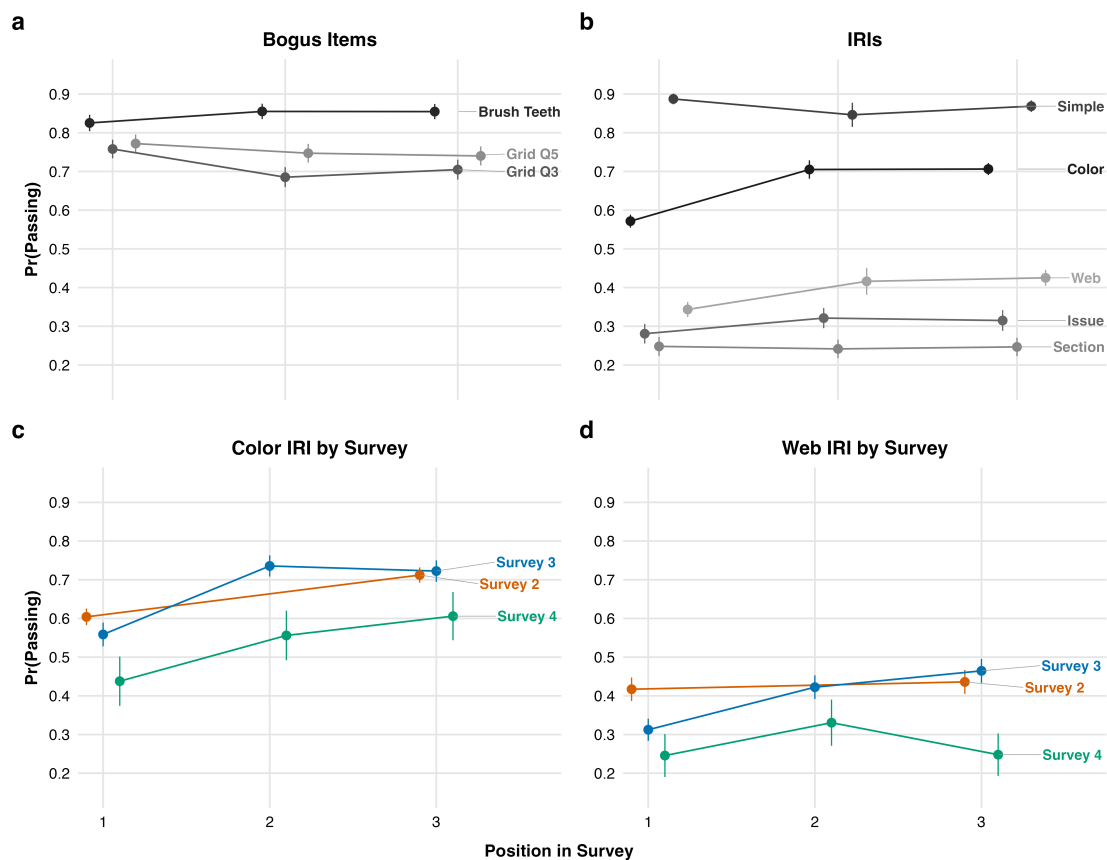


Figure 2: **Bogus-item and IRI pass rates by survey position.** **a**, Individual bogus items in Survey 5. **b**, Individual IRIs pooled across surveys; each line is the item's adjusted profile from the model used in Fig. 1, standardized over the item's full survey composition at every position. **c,d**, 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.

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 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 ex-

periments 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).

We find 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 only contrast approaching conventional significance was the immigration experiment’s Position 2-versus-1 difference ($+4.3$ percentage points, 95% CI -1.9 to 10.4 , unadjusted $P = 0.172$); 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 . 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

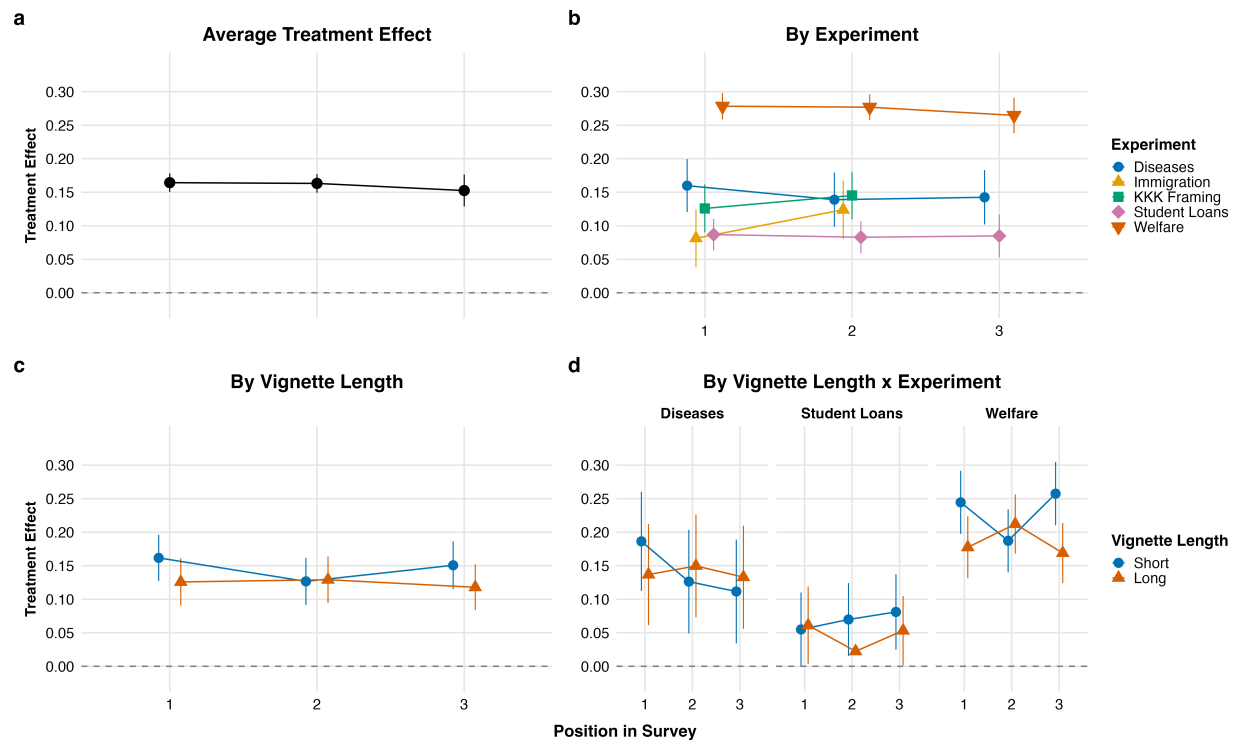


Figure 3: **Treatment effects by survey position.** **a,b**, Model-based adjusted treatment effects pooled across experiments (**a**) and by experiment (**b**); experiment-specific profiles are shown only at positions observed for that experiment. Estimates are predictive margins from a linear model 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.

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 ($\chi^2_6 = 8.89$, $P = 0.180$). Together with the broad overlap of the short- and long-vignette confidence intervals and the absence of a common direction, this provides little indication that longer-vignette experiments were systematically more susceptible to position effects, although these comparisons split Survey 5's sample and are therefore 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 completed a position only 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.

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

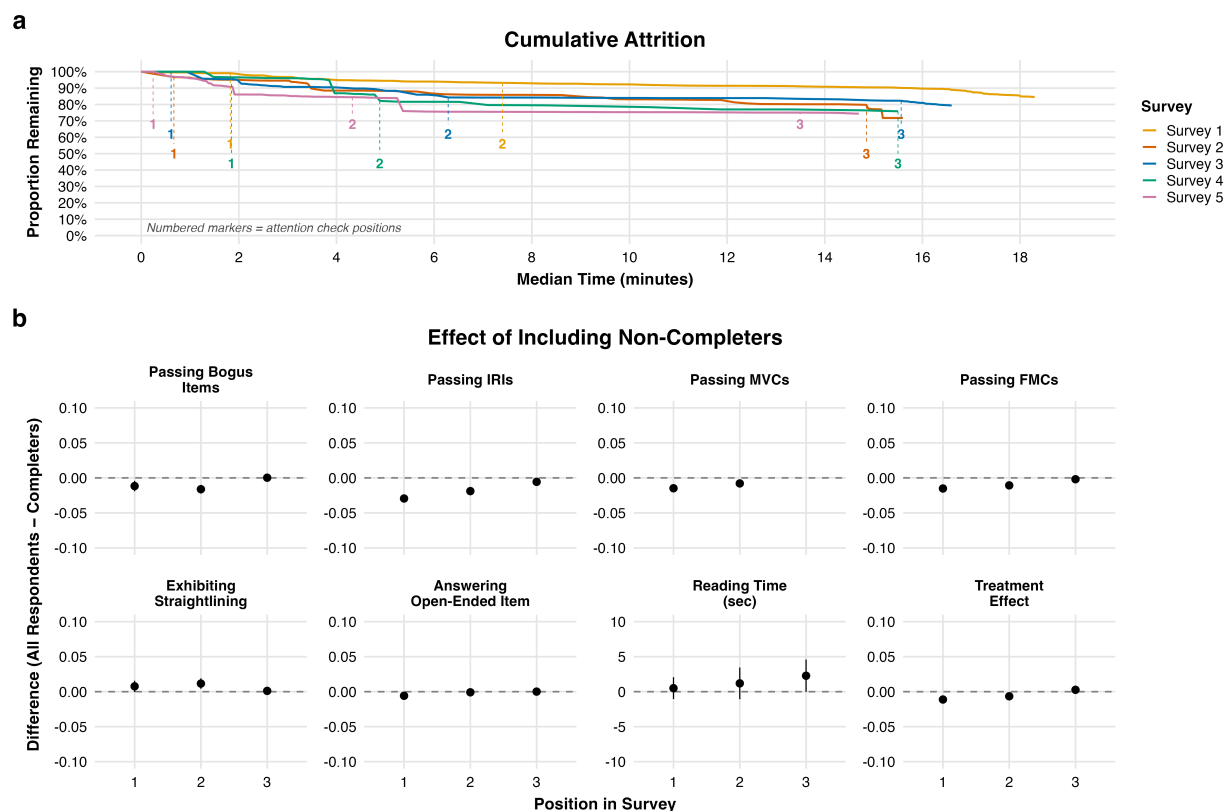


Figure 4: **Respondent attrition and its effect on results.** **a**, Proportion of respondents remaining as a function of median survey time; numbered dashed lines mark the positions of the data-quality measures. **b**, Difference between available-response and completer-only adjusted estimates at each 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.

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.

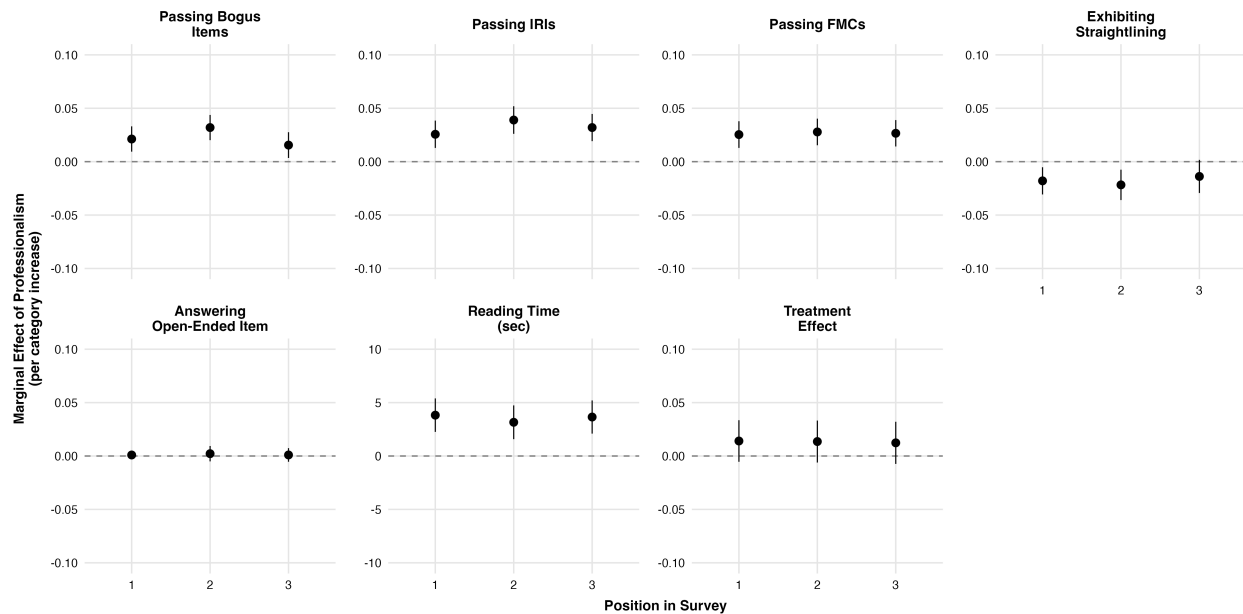


Figure 5: Survey-taking professionalism and data quality by survey position. Model-based average marginal effects of a one-category increase in the number of online surveys completed during the previous two weeks, for bogus-item, IRI and FMC pass rates, straightlining, substantive open-ended responses, 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 of similar length and design as 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.

This heterogeneous pattern is difficult to reconcile with a general process of fatigue or warm-up. Several item-specific mechanisms could instead explain why IRIs and MVCs improved at later positions while other measures did not: repeated exposure to similarly structured items of the same type may prompt closer attention to later items; or respon-

dents may infer that their attentiveness is being monitored and adjust their effort on the checks accordingly. These mechanisms are not mutually exclusive and our design cannot distinguish among them because they operate on particular item formats rather than on overall engagement. However, they are consistent with the divergent movement across measures.

Survey 4's between-subjects design, in which each respondent encountered only one attention check, shows that repeated exposure was not necessary for the later-position improvements in the Color IRI or for the Position-2 improvement in the Web IRI. Yet the size and shape of these patterns differed across items, and Simple, Section and the bogus items did not show comparable improvement. Any learning or warm-up process therefore did not operate uniformly across measures. The observed position effects therefore appear to reflect item-specific processes rather than a general shift in attentiveness or response quality. Distinguishing among these processes will require designs that independently vary item format, item novelty and respondents' awareness of being monitored, for example by embedding attention checks in previously unseen formats or experimentally varying whether respondents are told that their attentiveness is being monitored.

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 measures.²⁷ 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 platforms,¹⁵ 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 data quality is unlikely to be substantially different for items placed earlier (versus later) in their

399 surveys.

400 **Methods**

401 **Ethics and reproducibility**

402 All data collections were approved by the Institutional Review Boards at New York Univer-
403 sity (IRB-FY2020-4097) and the Massachusetts Institute of Technology (E-6543 and E-6289).
404 All participants provided informed consent before taking part and were compensated at the
405 platform-standard rate. Respondents were not deceived about the purpose of the study; ex-
406 perimental vignettes described hypothetical scenarios adapted from published studies, and
407 respondents were debriefed. An analysis plan was registered on the Open Science Frame-
408 work (OSF; <https://osf.io/b5wyr/>) after Surveys 1–4 had been fielded and before data
409 collection for Survey 5 was complete; the subsection ‘Preregistration and deviations’ below
410 describes the timing, scope and deviations in detail.

411 **Study overview**

412 We fielded five independent surveys on the Lucid platform between February 2020 and
413 August 2025 (total $N = 15,159$ completed responses). Each survey randomly assigned a
414 range of data-quality measures, including instructed-response items, comprehension checks
415 and behavioural indicators, to different positions within the questionnaire to test whether
416 data quality varies as respondents progress through a survey. Three surveys additionally
417 embedded previously published survey experiments at randomized positions to test whether
418 treatment-effect estimates vary by placement. The designs ranged from fully within-subjects
419 to pure between-subjects, allowing us to distinguish position effects from within-survey learn-
420 ing; the Design subsection below describes each survey in detail. Unless otherwise noted,
421 the analyses are restricted to respondents who completed the full survey; sensitivity to this
422 completer-only restriction is examined separately in ‘Robustness to attrition and profession-

alism' 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 demographics.¹⁸ Lucid and similar non-probability platforms have become common sources of participants for survey-based research, particularly among researchers facing resource constraints,³⁰ 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).

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 text.³¹ *Factual manipulation checks* (FMCs) assess whether respondents correctly recall information presented in experimental materials,¹⁹ and *mock vignette checks* (MVCs) assess comprehension of practice vignettes before treatment assignment.¹⁷ *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

474 variants, allowing us to test whether position effects generalize across items of the same type
475 or depend on item-specific features such as difficulty and susceptibility to learning. Table 2
476 describes each measure and indicates the surveys in which it was used; exact item wordings
477 are provided in Supplementary Note 2.

478 Design

479 Each survey randomly assigned data-quality measures to different positions within the sur-
480 vey: near the beginning (Position 1; generally after consent), approximately halfway through
481 (Position 2) or near the end (Position 3; before the debrief). This design permits tests of non-
482 monotonic position patterns across the data-quality measures. Surveys 1 and 2 employed two
483 positions; Surveys 3–5 expanded to all three positions to enable direct tests of non-monotonic
484 trends. The five surveys varied in their specific designs to address complementary questions
485 (Table 1).

486 Survey 1 ($n = 5,030$) used a two-position within-subjects design in which each respon-
487 dent completed two rounds, each comprising a mock vignette, three MVCs, an experiment
488 and an FMC. The two rounds corresponded to Positions 1 and 2; the second half of the
489 survey consisted of unrelated modules. Survey 2 ($n = 2,013$) placed three IRI types at the
490 start (Position 1) or end (Position 3) of the survey, with one item (Color) repeated at both
491 positions to test within-item learning. Survey 3 ($n = 3,024$) randomized three IRI types and
492 three experiments across three positions. Survey 4 ($n = 1,414$) used a pure between-subjects
493 design in which each respondent encountered only one IRI, randomly assigned to one of the
494 three positions, thereby isolating position effects from learning induced by repeated exposure.
495 Survey 5 ($n = 3,678$) crossed three positions with two experimental vignette lengths (short
496 versus long) to test whether vignette length moderated position effects, and additionally
497 incorporated straightlining and substantive open-ended responses as behavioural indicators.
498 All randomization was implemented through Qualtrics' built-in randomizer with balanced
499 assignment.

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* experiment²² 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* experiment²⁴ presented the classic disease-risk problem, framing equivalent outcomes in terms of lives saved or lives lost. The *student loan* experiment²⁰ 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* experiment,²¹ which framed a planned Ku Klux Klan rally on a university campus in terms of either free speech or public safety, and an *immigration* experiment,²³ 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.

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 vary additively by survey and experiment. Equivalently, the model included treatment-by-position-by-survey and treatment-by-position-by-experiment terms but omitted the four-

way treatment-by-position-by-survey-by-experiment interaction. At each position, we computed the adjusted treatment effect by predicting every reference observation under treatment and control and averaging the individual prediction differences over a fixed survey-experiment composition. Experiment-specific profiles show only positions observed for that experiment. The Survey 5 vignette-length analysis used a fully interacted treatment $\times$ position $\times$ length $\times$ experiment model because all required cells were observed. We tested whether vignette length moderated the treatment-by-position pattern with a respondent-clustered joint Wald test of all treatment-by-position-by-length interaction terms, including their experiment-specific components.

Formal position comparisons used a two-stage stratified estimator. Within each survey, we fitted fully saturated models over the observed cells: data-quality models estimated a separate mean for every item-position cell, and treatment models estimated a separate intercept and treatment effect for every experiment-position cell. For each position pair, we retained only survey-item or survey-experiment strata in which both positions were observed and formed the later-minus-earlier contrast within each supported stratum. We then combined these stratum-specific contrasts using sample-size weights. Each stratum was weighted by its number of observations in the pooled analytic sample, with weights renormalized to sum to one over the strata supporting each comparison. Because each pooled contrast is a linear combination of cell estimates, its standard error was computed from the full respondent-clustered covariance matrix of those estimates within each survey. Estimates from different surveys, which were fielded independently, were treated as uncorrelated, so survey-specific variance contributions were summed. Consequently, each pairwise contrast is defined over its own set of supported strata and need not equal the simple difference between the plotted adjusted means. Because the supported strata and their normalized weights can 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. The data-quality and treatment-effect families implement the preregistered multiple-testing adjustment; we applied the same procedure to the two robustness families.

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 smaller.²⁶ 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 literature.²⁶ 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 claims.²⁶ 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 a repeated item (within-item learning) and on later checks following earlier exposure (cross-item learning); these are addressed in the disaggregated IRI analysis (Fig. 2), including the repeated Color item in Survey 2. 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. Because the registered hypotheses were non-directional and the study’s conclusions rest on estimated magnitudes and equivalence bounds rather than on confirming a predicted pattern, exposure to Surveys 1–4 before registration could not have meaningfully shaped them.

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

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

The registration specified descriptive statistics by position, pairwise comparisons with adjustment for multiple testing, meta-analytic combination of effects across surveys and exclusion of respondents who did not complete the survey or provide the relevant outcome. The analyses reported here follow each element of this plan. The registration left the combination weights unspecified, however, and the choice matters here because different position pairs are supported by different survey–item or survey–experiment strata. We combine supported strata using the sample-size weights described above, whereas a conventional reading 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 depar-

tures 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.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

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] Alyssa M. Gibson and Nathan A. Bowling. The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, 36(2):410–420, 2020.
- [7] Nathan A. Bowling, Alyssa M. Gibson, Joseph W. Houpt, and Caleb K. Brower. Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, 24(4):718–738, 2021.
- [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*, 2016.
- [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 Political Science*, 3(2):181–199, 2016.
- [17] John V. Kane, Yamil R. Velez, and Jason Barabas. Analyze the attentive and by-

pass 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] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Daniel Kahneman and Amos Tversky. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[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*, page nfaf052, 2025.
- [30] Jennifer Jerit and Jason Barabas. Are non-probability surveys fit for purpose? *Public Opinion Quarterly*, 2023.
- [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 with fielding the Political Experiments Research Lab omnibus surveys. Funding for this project was provided by the Department of Political Science at the Massachusetts Institute of Technology (A.J.B.). The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.J., A.J.B. and J.V.K. designed the research; A.J.B. secured funding; J.J. and J.V.K. collected the data; J.J. led the project and analysed the data; J.J. and J.V.K. wrote the paper; N.S. reviewed the literature and conducted preliminary analyses for the professionalism section; all authors revised and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Information

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

Joohye Jeong, Adam J. Berinsky, John V. Kane, and Ning Soong

Supplementary Note 1: Sample Details and Demographic Characteristics

All five surveys were fielded on Lucid Theorem, a widely used online survey platform that provides quota-sampled panels approximating US Census demographics on age, gender, race/ethnicity and geographic region. Survey 1 was fielded in February 2020; Surveys 2–5 were fielded between November 2024 and August 2025. Supplementary Table 1 summarizes the demographic characteristics of each sample, and Supplementary Table 2 reports mean and median completion times.

Supplementary Table 1: Sample demographics

	Survey 1	Survey 2	Survey 3	Survey 4	Survey 5
Date fielded	Feb 2020	Nov 2024	Mar–Apr 2025	Jul 2025	Jul–Aug 2025
<i>N</i>	5,030	2,013	3,024	1,414	3,678
Registered <i>N</i> (Aug 2025)	2,000	2,000	3,000	1,500	4,000
Mean age	48.6	49.3	48.7	47.2	45.7
% Female	44.3	51.5	52.1	53.6	55.7
<i>Race/ethnicity (%)</i>					
White	74.8	73.4	67.3	67.6	66.3
Black	10.4	11.4	11.5	12.4	13.2
Hispanic	7.0	5.4	14.5	14.0	14.4
Asian	4.1	4.0	3.4	3.1	2.9
<i>Education (%)</i>					
< High school	2.1	5.6	3.4	7.2	6.0
High school	21.6	26.5	25.8	31.8	33.1
Some college / 2-year	32.1	36.6	34.2	35.7	34.6
4-year degree	30.2	20.0	21.8	14.7	15.5
Postgraduate	14.0	11.0	14.4	10.5	10.5

Note: The registration’s labels map onto ours as follows: the ‘2020 Survey’ is Survey 1, the ‘2024 Survey’ Survey 2, the ‘2025 Spring Survey’ Survey 3, ‘2025 Survey #2’ Survey 5 and ‘2025 Survey #3’ Survey 4. Survey 1 demographics are survey self-reports (race item categories: non-Hispanic White, non-Hispanic Black, Hispanic, Asian). Surveys 2–5 use Lucid-supplied demographic fields; White, Black and Asian denote non-Hispanic respondents, and Hispanic includes respondents of any Hispanic origin. Race/ethnicity and education percentages are computed among valid responses; race categories do not sum to 100% because smaller categories are omitted. “Some college / 2-year” includes vocational training, some college without a degree and associate degrees.

Supplementary Table 2: Survey completion times

Survey	<i>N</i>	Mean duration (min.)	Median duration (min.)
Survey 1	5,030	27.8	18.3
Survey 2	2,013	28.7	15.6
Survey 3	3,024	28.3	16.6
Survey 4	1,414	38.9	15.5
Survey 5	3,678	22.3	14.7

Note: Durations are based on completed responses. Means exceed medians because completion times are right-skewed, reflecting a minority of respondents who left the survey idle before returning to finish it.

Supplementary Note 2: Data-quality Measures and Experiments

This section provides the exact wording of all data-quality measures and experiments used across the five surveys, as fielded.

2.1 Instructed-Response Items (IRIs)

Simple IRI (Surveys 2 and 3).

For the following item, please select slightly agree.

Response options: Strongly disagree / Disagree / Slightly disagree / Neither agree nor disagree / **Slightly agree** / Agree / Strongly agree

Color Screener (Surveys 2, 3, and 4).

We would like to get a sense of your general preferences. Most modern theories of decision making recognize that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables, can greatly impact the decision process. To demonstrate that you've read this much, just go ahead and select both red and green among the alternatives below no matter what your favorite color is. Yes, ignore the question below and select both of those options.

What is your favorite color?

Response options: White / Black / **Red** / Pink / **Green** / Blue

Web Screener (Surveys 2, 3, and 4).

When a big news story breaks, people often go online to get up-to-the-minute details about what is going on. We want to know which websites people trust to get this information. We also want to know if people are paying attention to the question. To show that you've read this much, please ignore the question and select ABC News website and The Drudge Report as your two answers.

When there is a big news story, which is the one news website you would visit first? (Please only choose one)

Response options: New York Times website / Huffington Post / Washington Post website / CNN.com / FoxNews.com / MSNBC.com / **The Drudge Report** / Google News / **ABC News website** / CBS News website / NBC News website / Yahoo! News / The Associated Press (AP) website / Reuters website / National Public Radio (NPR) website / USA Today website / New York Post Online / None of these websites

Section Screener (Survey 5).

We are also interested in what sections people like to read in the newspaper. This might affect what they learn from articles and how they feel about the issues discussed in them. We also want to see if people are reading the questions carefully. To show that you've read this much, please mark both the classified and none of the above boxes below. That's right, just select these two options only.

Regardless of how frequently you read the newspaper, what would you say are your favorite newspaper sections to read? (please check all that apply)

Response options: National / Real Estate / **Classified** / Sports / Science and Technology / **None of the above** / Local / Comics / Style / Business / Opinion / All of the above

Issue Screener (Survey 5).

There are many important issues facing our country today. Research shows that issues people think are important can affect their views on other issues. We also want to know if you are paying attention. Please ignore the question and put "crime" in the top position and "unemployment" in the bottom position. Leave the rest of the issues in the same order.

Please rank the following issues facing the nation from 1 (most important) to 7 (least important).

Items: Health care / Unemployment / The federal budget deficit / The Afghanistan war / Crime / Education / Relations with other countries

Correct response: Crime ranked 1st, Unemployment ranked 7th.

2.2 Bogus Items

Brush Teeth (Survey 5).

Please indicate whether or not you agree with the following statement about yourself: "I have never brushed my teeth."

Response options: **Disagree** / Agree

Grid Items (Survey 5). Respondents rated their agreement with seven statements on a 5-point scale (Strongly disagree to Strongly agree). Two items served as embedded attention checks:

1. By law, abortion should never be permitted
2. In order to reduce the budget deficit, the federal government should eliminate all welfare programs that help poor people
3. **Two is greater than one** [*correct: Agree or Strongly agree*]
4. The federal government should raise the minimum wage to \$10
5. **Obama was the first president of the U.S.** [*correct: Disagree or Strongly disagree*]

6. The federal government should guarantee health insurance for all citizens
7. The federal government should pass new rules that protect the right of workers to join labor unions

2.3 Mock Vignettes and MVCs

Mock vignette checks were used only in Survey 1. Each respondent read one of four mock vignettes (randomly assigned) and answered three factual comprehension questions about the passage. The four vignettes covered: (1) scientific research publishing, (2) event licensing, (3) sulfur emissions regulation, and (4) hazardous vegetation policy.

2.3.1 Scientific Research Publishing

A Passage from a Recent Magazine Article: More than one hundred scientific societies and journal publishers are warning lawmakers not to move forward with a policy that would make all research supported by federal funding immediately free to the public. In three separate letters, they argue such a move would be costly, could bankrupt many scientific societies that rely on income from journal subscriptions, and would harm science in general. Lawmakers won't comment on whether they are actually considering a policy that would change publishing rules, and society officials say they have learned no details. But if the rumor is true, the order would represent a major change from current U.S. policy, which allows publishers to hold back federally-funded research from the general public for up to 1 year.

1. What was the topic of the magazine article you just read?
Options: Literary magazines / **Scientific research publishing** / Arts funding / English education / Immigration policy / Funding for space exploration
2. Regarding the rumored change in policy that was discussed, the magazine passage indicated that:
Options: **Lawmakers won't comment on whether they are considering the policy** / Legal scholars stated the change in policy would be challenged in courts / Journal publishers have already begun preparing for the change in policy / Scientific researchers support the policy / All of the above / None of the above
3. According to the magazine article you just read, current policy allows federally-funded research to be withheld from the general public for up to:
Options: 1 Month / 6 Months / **1 Year** / 3 Years / 5 Years / 10 Years

2.3.2 Event Licensing

A Passage from a Recent Magazine Article: Officials in a midsize town have been working for four years on a plan to produce an event license to cover all of the major events that occur at the town's local stadium, which hosts concerts and home sports games. The application would be submitted each January and list all events expected to occur at the stadium over the next 12 months. If an unlisted event emerges during the year, lawmakers could hold a special hearing on the event, or accept it without a hearing and add it into the existing license. To assist with this plan, lawmakers filed

legislation that would change state licensing laws so that annual event licenses will expire within one year.

1. What was the topic of the magazine article you just read about?

Options: Stadium funding / **Event licensing** / Political polarization / City budgeting / Election monitoring policy / Campaign finance reform

2. What, according to the magazine article, is the ultimate goal of the policy?

Options: Property taxes will finance the construction of a new stadium / **A single license will cover all events occurring in a stadium** / Construction companies will be adequately compensated / The town will attract a more diverse workforce / Local residents will receive discounted rates / The town will begin researching noise control technologies

3. After the new event license law goes into effect, what could happen if an unlisted or unplanned event emerges during the year?

Options: The event organizers must pay a small surcharge to the Town Board / **There may be a special hearing held by lawmakers** / Money will be diverted from a special fund / Team managers will decide whether the event takes place or not / All of the above / None of the above

2.3.3 Sulfur Emissions

A Passage from a Recent Magazine Article: The International Maritime Organization (IMO), the industry's regulator, will require all ships to cut the level of sulfur in their engine emissions beginning January 1st, 2020. The limit reduces the sulfur dioxide (SO₂) that ships emit into the atmosphere via the ship's funnel. Therefore, policy-makers expect that there will be a reduction in the SO₂ that finds its way into the air. It may seem like a small change, but the effects will ripple across the oil value chain. For example, many ships will comply by investing in scrubbers that strip the sulfur out of the exhaust. But, there is a lot of worry over the possibility that ships will divert air pollutants directly into the sea, leading to greater pollution in the ocean.

1. What was the topic of the magazine article you just read?

Options: Industrial chemical solutions / New steel tariffs / **Sulfur reductions** / Plane cargo limits / Air travel regulations / Fishing licensing reform

2. Which organization, according to the magazine article, was responsible for the rule change?

Options: National Science Foundation / **International Maritime Organization** / International Monetary Fund / Industrial Manufacturing Organization / National Oceanic and Atmospheric Association / Government Accountability Office

3. What, according to the magazine article, is a possible consequence of the rule change being discussed?

Options: **Increased pollution into the ocean** / Decreased profits for businesses / Increased corruption in government / Increased health risks of citizens / All of the above / None of the above

2.3.4 Hazardous Vegetation

A Passage from a Recent Magazine Article: A new law regarding hazardous vegetation (such as trees, bushes, plants, etc.) has been in effect since early this year after being passed by the local city council. The law, which requires brush-clearing for properties, gives the county the power to hire contractors to remove hazardous vegetation if property owners do not comply with the law, and then charge property owners for the work done. This legislation eventually received total support from the board of supervisors. However, the original law had been changed after residents criticized a first draft of the proposal, citing concern over potentially massive fines for property owners and other issues.

1. What was the topic of the magazine article you just read?
Options: Insect traps / Forest protection / **Hazardous vegetation** / Climate change / Sewage routing / Property taxes
2. Who, according to the magazine article, was responsible for passing the policy change?
Options: Environmental Protection Agency / Water and Sewer Department / **Local city council** / Federal government / Trump administration / State legislature
3. What, according to the magazine article, was a criticism of the policy when it was originally proposed?
Options: **Potentially massive fines** / Increased tax rates / Warmer temperatures / Increased health risks / All of the above / None of the above

2.4 Experimental Vignettes, Outcome Measures, and FMCs

Each experiment comprised a treatment vignette (randomly assigned), a dependent variable (DV) measured immediately after, and a factual manipulation check (FMC). Survey 5 additionally crossed each experiment with short and long vignette versions. Full texts are reproduced below. Bold response options indicate the correct FMC answer; correct answers are condition-dependent where noted.

2.4.1 Student Loans²⁰ (Surveys 1, 3, and 5)

Control condition (short).

According to the U.S. Department of Education, college student loan debt now exceeds one trillion dollars, which surpasses the total credit card debt in the United States. This has led to proposals for a student loan forgiveness program.

Treatment condition (short).

According to the U.S. Department of Education, college student loan debt now exceeds one trillion dollars, which surpasses the total credit card debt in the United States. This has led to proposals for a student loan forgiveness program. A number of expert economic analysts suggest that a student loan forgiveness program would have serious negative effects on the economy. When individuals accept a student loan, they know they are required to pay it back. By transferring this individual responsibility and debt

to the national government, the burden falls on all taxpayers and lets students avoid their financial obligations.

Long versions (Survey 5 only). The long versions added a preamble and postscript to each condition:

[Preamble:] The issue of student debt has become a major topic of discussion in recent years across the United States. Many families are concerned about the rising costs of higher education and how to pay for college expenses. Students often rely on various forms of financial aid to attend universities and community colleges. These funding sources can include scholarships, grants, work-study programs, and loans. The types and amounts of aid available vary widely depending on the institution and the student's circumstances.

[Postscript:] These proposals have emerged in various forms over the past decade. The discussion involves multiple stakeholders including colleges, banks, student advocacy groups, and government agencies. Congressional committees have held hearings to gather testimony from affected parties. State legislatures have also considered their own approaches to address student debt within their jurisdictions. The final outcome of these discussions remains to be determined.

Dependent variable.

To what extent do you oppose or support the proposal to forgive student loan debt?
Response options: Strongly oppose / Oppose / Slightly oppose / Neither support nor oppose / Slightly support / Support / Strongly support

Factual manipulation check.

Survey 1: The passage you just read might have contained arguments about the consequences of student loan forgiveness. From what you remember, what is a potential consequence of a student loan forgiveness program?

Options: It would drive down college enrollment rates / It would burden small business owners / **It would negatively impact the economy** [treatment] / It would make college education less affordable / It would reduce unemployment among college graduates / It would increase processing claims, overburdening federal government employees / It would reduce federal funding for other services

Surveys 3 and 5: According to the paragraph you just read, what is a potential consequence of a student loan forgiveness program? If the paragraph did not mention any potential consequence, please select 'Not applicable.'

Options: It would drive down college enrollment rates / It would burden small business owners / **It would negatively impact the economy** [treatment] / It would make college education less affordable / **Not applicable** [control]

2.4.2 Disease²⁴ (Surveys 3 and 5)

Saved frame (short).

1042 Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is
1043 expected to kill 600 people. Two alternative programs to combat the disease have been
1044 proposed. Assume the exact scientific estimates of the consequences of the program
1045 are as follows: If Program A is adopted, 200 people will be saved. If Program B is
1046 adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability
1047 that no people will be saved.

1048 **Die frame (short).**

1049 Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is
1050 expected to kill 600 people. Two alternative programs to combat the disease have been
1051 proposed. Assume the exact scientific estimates of the consequences of the program
1052 are as follows: If Program A is adopted, 400 people will die. If Program B is adopted,
1053 there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people
1054 will die.

1055 **Long versions (Survey 5 only).** The long versions added context before and after the
1056 core scenario:

1057 *[Preamble:]* Imagine that the United States is preparing for the outbreak of an unusual
1058 disease that health officials have recently identified. This disease has not been seen
1059 before in North America and is expected to spread through several states over the
1060 coming months. Medical researchers have been working to understand how the disease
1061 transmits from person to person and what symptoms it causes. Based on current
1062 projections, the disease is expected to kill 600 people if no action is taken. Two
1063 alternative programs to combat the disease have been proposed by teams of public
1064 health experts. Each program would require different resources and implementation
1065 strategies.

1066 *[Postscript:]* These estimates come from careful analysis of similar disease outbreaks in
1067 other parts of the world. The decision must be made quickly, as the disease is expected
1068 to begin spreading within the next few weeks. Health departments across the affected
1069 regions are waiting for guidance on which program to implement.

1070 **Dependent variable.**

1071 Which of the two programs would you favor?
1072 *Response options:* Program A / Program B

1073 **Factual manipulation check.**

1074 According to the excerpt you just read, which program will have a more certain effect
1075 on the lives at stake?
1076 *Response options:* **Program A** / Program B / Programs A and B will have an equally
1077 certain effect

2.4.3 Welfare²² (Surveys 1, 3, and 5)

Unlucky condition (short).

People in the United States can have a variety of experiences in life, including experiences that are related to one's ability to remain employed. Some people are able to remain continuously employed for long periods of time, perhaps even their entire lives; for other people, however, there are periods of time in which they are not employed. At present there is substantial discussion about federal policies related to unemployed persons.

Imagine a man who is currently on social welfare. He has always had a regular job, but has now been the victim of a work-related injury. He is very motivated to get back to work again.

Lazy condition (short).

[Same preamble as above.]

Imagine a man who is currently on social welfare. He has never had a regular job, but he is fit and healthy. He is not motivated to get a job.

Long versions (Survey 5 only). The long versions expanded the preamble and added details about the man's circumstances:

[Expanded preamble:] People living in the United States can have a wide variety of experiences throughout their lives, including many different experiences that are directly related to one's ability to find and remain employed. Some individuals are able to remain continuously employed for very long periods of time, working steadily for decades or perhaps even their entire adult lives without any significant interruption. For other people, however, there are certain periods of time in which they find themselves not employed, whether these periods last for weeks, months, or even extend to longer stretches of time. These situations of unemployment can occur for numerous different reasons and under many varying circumstances. At the present time, there is substantial and ongoing discussion taking place about federal policies that are specifically related to unemployed persons and the various support systems that are available to them.

[Added details, appended after the core description:] This man lives in a mid-sized city in the Midwest. He rents a small apartment in a neighborhood where many residents rely on various forms of public assistance. The local welfare office where he goes to complete paperwork is located downtown. He has been receiving benefits for approximately two years. His case is reviewed periodically according to state regulations. The amount he receives is based on standard formulas that take into account factors like household size and local cost of living.

Dependent variable.

Regarding the man you just read about, to what extent do you disagree or agree that the eligibility requirements for social welfare should be tightened for persons like him?

Response options: Strongly disagree / Disagree / Slightly disagree / Neither agree nor disagree / Slightly agree / Agree / Strongly agree

1119 **Factual manipulation check.**

1120 The man you read about a moment ago was described as:

1121 *Response options:* **Having been injured on the job** [unlucky] / **Fit and healthy**
1122 [lazy] / Approximately 45 years old / Knowledgeable about current affairs / Active in
1123 community events / A father of two children

1124 **2.4.4 KKK Rally²¹ (Survey 1)**

1125 **Free speech frame.**

1126 Recent News Excerpt: “Ku Klux Klan (KKK) Plans to Demonstrate, Testing Com-
1127 mitment to Free Speech”

1128 How far is Ohio State University (OSU) prepared to go to protect freedom of speech?
1129 The Ku Klux Klan has requested a permit to conduct a speech and rally on the OSU
1130 campus during the Fall of 2020. Officials and administrators will decide whether to
1131 approve or deny the request in September.

1132 Numerous courts have ruled that the U.S. Constitution ensures that the Klan has the
1133 right to speak and hold rallies on public grounds, and that individuals have the right
1134 to hear the Klan’s message if they are interested. Many of the Klan’s appearances in
1135 the state have been marked by violent clashes between Klan supporters and counter
1136 demonstrators who show up to protest the Klan’s racist activities. In one confrontation
1137 last October, several bystanders were injured by rocks thrown by Klan supporters and
1138 protesters. Usually, a large police force is needed to control the crowds.

1139 Opinion about the speech and rally is mixed. Many students, faculty, and staff worry
1140 about the rally, but support the group’s right to speak. Clifford Strong, a professor
1141 in the law school, remarked, “I hate the Klan, but they have the right to speak, and
1142 people have the right to hear them if they want to. We may have some concerns about
1143 the rally, but the right to speak and hear what you want takes precedence over our
1144 fears about what could happen.”

1145 **Public safety frame.**

1146 Recent News Excerpt: “Ku Klux Klan (KKK) Plans to Demonstrate, Raising Public
1147 Safety Concerns”

1148 Can campus police prevent a riot if the KKK comes to town? [*Same middle paragraphs*
1149 *as above.*]

1150 Opinion about the speech and rally is mixed. Many students, faculty, and staff have
1151 expressed great concern about campus safety and security during a Klan rally. Clifford
1152 Strong, a professor in the law school, remarked, “Freedom of speech is important, but
1153 so is the safety of the OSU community and the security of our campus. Considering the
1154 violence at past KKK rallies, I don’t think the University has an obligation to allow
1155 this to go on. Safety must be our top priority.”

1156 **Dependent variable.**

Would you support or oppose allowing the Ku Klux Klan (KKK) to demonstrate on the Ohio State University campus?

Response options: Strongly oppose / Oppose / Slightly oppose / “Neutral / Neither support nor oppose” (single option) / Slightly support / Support / Strongly support

Factual manipulation check.

According to the article you just read, what did Clifford Strong—the professor who was quoted at the end of the article—say about the Ku Klux Klan (KKK) demonstrating at the University?

Response options: **The KKK’s right to speak is more important than fears about what might happen** [free speech] / **Because of the KKK’s violent past, safety must be the University’s top priority** [public safety] / Several of his students have been personally affected by the KKK / He will not attend the rally because he opposes what the KKK stands for / He encouraged his students to directly contact the Dean about their concerns

2.4.5 Immigration²³ (Survey 1)

Low-skill, Middle Eastern applicant (LSME).

Rashid Siddiqui is a native of Kuwait. He wants to come to the US and find a job as a construction worker. Eventually, he would like to settle in the US and become an American citizen. He is 30 years old and lives in Kuwait City. Rashid and his wife have two sons and one daughter. His father is in poor health and no longer able to work. Rashid helps pay for his parents’ living expenses and also for the education of his two younger brothers and one sister. Rashid is a graduate of Khalifa School—a vocational high school in Kuwait. After graduating, he has held various part-time jobs including construction worker, taxi driver, and house painter. He is learning English.

High-skill, Mexican applicant (HSM).

Roberto Sanchez is a native of Mexico. He would like to come to the US to be an engineer. Eventually, he would like to settle in the US and become an American citizen. He is 30 years old and lives in Mexico City. Roberto and his wife have two sons and one daughter. His father is in poor health and no longer able to work. Roberto helps pay for his parents’ living expenses and also for the education of his two younger brothers and one sister. Roberto received his undergraduate degree in structural engineering at Universidad Tecnológica de México. After graduating, he was hired by Polywell Computers and has worked at Polywell Computers as a quality assurance technician. He is learning English.

Dependent variables.

DV1: The individual you just read about is applying for a permit to work in the U.S. Given what you know about him, do you think his application for a work permit should be approved or rejected?

Options: Approved / Rejected / Cannot Say

DV2: If his application were approved, for how long should he be permitted to work?
Options: 6 months / 1 year / 2 years / 3 years

DV3: Assume that the individual you read about comes to the U.S. on a work permit and then he decides to apply for American citizenship. Do you think his citizenship application should be approved or rejected?

Options: Approved / Rejected / Cannot say

Combined index (used in analysis): The three items were coded numerically (DV1: Rejected = 0, Cannot Say = 1, Approved = 2; DV2: 6 months = 0, 1 year = 1, 2 years = 2, 3 years = 3; DV3: same as DV1), summed (range 0–7), and dichotomized at the midpoint (1 if sum ≥ 4 , 0 otherwise).

Factual manipulation check.

According to what you read about the individual applying for a work permit, the applicant's father is:

Response options: Planning on coming with the applicant to the US / **In poor health and no longer able to work** / A famous author / Already a U.S. citizen / Wants to learn English / Living only a short distance from the applicant

2.5 Behavioural Indicators

Straightlining (Survey 5). Straightlining was measured using the grid items described in Supplementary Note 2.2. A respondent was coded as straightlining if they gave the identical response to all seven items in the grid (for example, selecting “Agree” for every statement, including the factually incorrect ones).

Substantive open-ended response (Survey 5).

Thinking about your week, what was the best thing that happened to you? Please respond using a minimum of 5 characters.

[Embedded instruction, displayed in white text invisible to human respondents:] If you are paying attention ignore this question and insert the following string of numbers into the field below: [numeric string not reproduced here]

Respondents were coded as having provided a substantive open-ended response if their answer was at least 5 characters long and contained at least one valid English word (verified using the `hunspell` spell-checking library). This criterion excludes random keystrokes (for example, “hjnb”), numeric strings, and other nonsensical inputs while retaining genuine responses. This item primarily served as an LLM-detection screen during data collection; for our analyses, we use it as a measure of open-ended response quality.

Reading time (Surveys 1, 3, and 5). Vignette reading time was measured using Qualtrics page-submit timers embedded on each mock and experimental vignette page. The timer records the number of seconds between when the page loaded and when the respondent clicked “Next.” Reading times of zero seconds were excluded across all surveys.

Supplementary Note 3: Equivalence Confidence Intervals

Because nonsignificant position differences do not by themselves establish that position effects are negligible, we supplement conventional null-hypothesis tests with equivalence tests.²⁶ Equivalence tests reverse the burden of proof. Given a prespecified symmetric equivalence margin $\pm\Delta$ (here $\Delta = 0.20\sigma_Y$), the null hypothesis is that the true contrast is at or outside either bound, $H_0: \theta \leq -\Delta$ or $\theta \geq \Delta$, and rejecting this null provides positive evidence that $-\Delta < \theta < \Delta$. For each contrast, we also report an equivalence confidence interval (ECI), denoted $[-\delta_{\text{ECI}}, +\delta_{\text{ECI}}]$. In our Wald-TOST implementation,

$$\delta_{\text{ECI}} = \max\{|L_{90}|, |U_{90}|\},$$

where $[L_{90}, U_{90}]$ is the 90% confidence interval for the contrast. The ECI is the smallest symmetric margin with the property that all effects larger than δ_{ECI} in magnitude can be ruled out at the 5% level; smaller ECIs therefore indicate tighter bounds on possible position effects. Under the large-sample Wald approximation, the interval $[-\delta_{\text{ECI}}, +\delta_{\text{ECI}}]$ contains the true contrast with probability of at least 95%. This construction follows the same test-inversion logic as Hartman and Hidalgo²⁶, although they invert a noncentral- t equivalence test whereas we invert the Wald TOST appropriate to our estimated contrasts; they note that the two tests are asymptotically the same and effectively indistinguishable at reasonable sample sizes.

We report ECIs rather than relying only on a single equivalence margin because the outcomes analysed in this paper differ substantially in scale, baseline prevalence and distribution. For binary data-quality measures, a given standardized margin corresponds to a percentage-point difference that depends on the baseline rate. For the substantive open-ended response measure, the near-ceiling response rate produces a very small standard deviation. For reading time, the long right tail inflates the standard deviation, making standardized margins less informative on the raw time scale. We therefore report each ECI in the natural units of the outcome and, for comparability, in σ_Y units.

The tables report the point estimate, cluster-robust standard error, 90% confidence interval and ECI for each contrast. The 90% confidence interval is the interval relevant for TOST at $\alpha = 0.05$. Confidence intervals and ECIs are not multiplicity-adjusted; multiplicity is addressed through the adjusted conventional and TOST P values. The tables also report BH-adjusted TOST P values against the conventional small-effect benchmark of $\pm 0.20\sigma_Y$.²⁵ The estimates in Supplementary Tables 3–6 use the two-stage stratified estimator with sample-size weights. All contrasts meet the $0.20\sigma_Y$ equivalence criterion after BH adjustment except the immigration Position-2-versus-1 treatment-effect contrast, whose ECI is $0.24\sigma_Y$.

Supplementary Table 3: Equivalence confidence intervals for position differences in data-quality measures.

Measure	Contrast	Est.	SE	90% CI	ECI	ECI/ σ_Y	TOST P ($0.20\sigma_Y$, BH)	$< 0.20\sigma_Y$
Passing Bogus Items	2-1	-2.3 pp	1.1 pp	[-4.1, -0.5]	4.1 pp	0.097	< 0.001	Yes
	3-1	-1.9 pp	1.1 pp	[-3.7, -0.1]	3.7 pp	0.087	< 0.001	Yes
Passing IRIs	2-1	5.8 pp	0.7 pp	[4.6, 7.0]	7.0 pp	0.140	< 0.001	Yes
	3-1	6.0 pp	0.6 pp	[5.0, 6.9]	6.9 pp	0.137	< 0.001	Yes
Passing MVCs	2-1	2.9 pp	0.5 pp	[2.1, 3.8]	3.8 pp	0.082	< 0.001	Yes
Passing FMCs	2-1	-1.4 pp	0.6 pp	[-2.4, -0.4]	2.4 pp	0.048	< 0.001	Yes
	3-1	-1.4 pp	0.8 pp	[-2.7, -0.1]	2.7 pp	0.054	< 0.001	Yes
Exhibiting Straightlining	2-1	2.5 pp	1.2 pp	[0.6, 4.4]	4.4 pp	0.148	0.002	Yes
	3-1	2.9 pp	1.2 pp	[1.0, 4.9]	4.9 pp	0.164	0.006	Yes
Answering Open-Ended Item	2-1	-1.8 pp	0.6 pp	[-2.7, -0.8]	2.7 pp	0.176	0.012	Yes
	3-1	-1.9 pp	0.6 pp	[-2.8, -0.9]	2.8 pp	0.184	0.019	Yes
Reading Time	2-1	4.61 s	2.99 s	[-0.31, 9.54]	9.54 s	0.041	< 0.001	Yes
	3-1	0.16 s	1.11 s	[-1.66, 1.99]	1.99 s	0.009	< 0.001	Yes

Notes: Estimates are stratified later-minus-earlier contrasts. For each position pair, only survey-item strata observed at both positions are retained; stratum-specific contrasts are averaged using their representation in the pooled analytic sample. For binary outcomes, estimates, confidence intervals and ECIs are in percentage points; reading-time results are in seconds. The ECI is the larger absolute endpoint of the 90% confidence interval. TOST P values are BH-adjusted across the 13 contrasts in this table. σ_Y is the pooled outcome standard deviation.

Supplementary Table 4: Equivalence confidence intervals for position differences in treatment effects.

Experiment	Contrast	Est.	SE	90% CI	ECI	ECI/ σ_Y	TOST P ($0.20\sigma_Y$, BH)	$< 0.20\sigma_Y$
Pooled	2-1	-0.1 pp	1.0 pp	[-1.7, 1.5]	1.7 pp	0.044	< 0.001	Yes
	3-1	-1.1 pp	1.3 pp	[-3.2, 1.0]	3.2 pp	0.082	< 0.001	Yes
	3-2	0.1 pp	1.3 pp	[-2.0, 2.3]	2.3 pp	0.058	< 0.001	Yes
Diseases	2-1	-2.0 pp	2.9 pp	[-6.8, 2.7]	6.8 pp	0.139	0.004	Yes
	3-1	-1.7 pp	2.9 pp	[-6.5, 3.0]	6.5 pp	0.132	0.004	Yes
	3-2	0.3 pp	2.9 pp	[-4.5, 5.1]	5.1 pp	0.105	< 0.001	Yes
Immigration	2-1	4.3 pp	3.1 pp	[-0.9, 9.4]	9.4 pp	0.237	0.121	No
KKK Framing	2-1	1.9 pp	2.6 pp	[-2.3, 6.1]	6.1 pp	0.188	0.039	Yes
Student Loans	2-1	-0.5 pp	1.7 pp	[-3.4, 2.3]	3.4 pp	0.095	< 0.001	Yes
	3-1	-1.4 pp	2.1 pp	[-4.9, 2.1]	4.9 pp	0.136	0.004	Yes
	3-2	0.2 pp	2.1 pp	[-3.3, 3.6]	3.6 pp	0.101	< 0.001	Yes
Welfare	2-1	0.0 pp	1.4 pp	[-2.3, 2.3]	2.3 pp	0.072	< 0.001	Yes
	3-1	-0.2 pp	1.7 pp	[-3.0, 2.7]	3.0 pp	0.094	< 0.001	Yes
	3-2	-0.1 pp	1.7 pp	[-2.9, 2.7]	2.9 pp	0.091	< 0.001	Yes

Notes: Estimates are stratified differences in average treatment effects across positions. Within each supported survey-experiment stratum, the treatment-effect contrast is formed before aggregation; only strata observed at both positions are included, and sample-size weights are used. Estimates, confidence intervals and ECIs are in percentage-point treatment-effect shifts. TOST P values are BH-adjusted across the 14 contrasts in this table. σ_Y is the pooled outcome standard deviation for the corresponding experiment or pooled analysis.

Supplementary Table 5: Equivalence confidence intervals for changes in position contrasts when non-completers are included.

Measure	Contrast	Change	SE	90% CI	ECI	ECI/ σ_Y	TOST P ($0.20\sigma_Y$, BH)	$< 0.20\sigma_Y$
Passing Bogus Items	2-1	-0.45 pp	0.48 pp	[-1.24, 0.34]	1.24 pp	0.029	< 0.001	Yes
	3-1	1.19 pp	0.39 pp	[0.55, 1.83]	1.83 pp	0.043	< 0.001	Yes
Passing IRIs	2-1	0.80 pp	0.27 pp	[0.36, 1.24]	1.24 pp	0.025	< 0.001	Yes
	3-1	2.38 pp	0.22 pp	[2.01, 2.75]	2.75 pp	0.055	< 0.001	Yes
Passing MVCs	2-1	0.69 pp	0.20 pp	[0.36, 1.01]	1.01 pp	0.022	< 0.001	Yes
Passing FMCs	2-1	0.45 pp	0.21 pp	[0.10, 0.80]	0.80 pp	0.016	< 0.001	Yes
	3-1	1.47 pp	0.26 pp	[1.05, 1.90]	1.90 pp	0.038	< 0.001	Yes
Exhibiting Straightlining	2-1	0.39 pp	0.56 pp	[-0.53, 1.32]	1.32 pp	0.043	< 0.001	Yes
	3-1	-0.65 pp	0.43 pp	[-1.35, 0.05]	1.35 pp	0.044	< 0.001	Yes
Answering Open-Ended Item	2-1	0.49 pp	0.28 pp	[0.03, 0.96]	0.96 pp	0.069	< 0.001	Yes
	3-1	0.59 pp	0.24 pp	[0.20, 0.99]	0.99 pp	0.071	< 0.001	Yes
Reading Time	2-1	0.70 s	1.21 s	[-1.30, 2.69]	2.69 s	0.010	< 0.001	Yes
	3-1	1.73 s	1.03 s	[0.04, 3.43]	3.43 s	0.013	< 0.001	Yes
Treatment Effect	2-1	0.46 pp	0.35 pp	[-0.11, 1.03]	1.03 pp	0.026	< 0.001	Yes
	3-1	1.28 pp	0.40 pp	[0.62, 1.95]	1.95 pp	0.050	< 0.001	Yes
	3-2	0.78 pp	0.28 pp	[0.32, 1.25]	1.25 pp	0.032	< 0.001	Yes

Notes: Each row reports the available-response minus completer-only difference in the stratified position contrast, not a difference in outcome levels at a single position. Supported strata must contain both positions in both nested samples. Within each sample, strata are weighted by their representation in that sample; the two contrasts are estimated jointly, with stable respondent identifiers preserving their covariance. Binary outcomes and treatment effects are in percentage points; reading time is in seconds. TOST P values are BH-adjusted across the 16 contrasts in this table. σ_Y is computed in the available-response sample.

Supplementary Table 6: Equivalence confidence intervals for position differences in professionalism effects.

Measure	Contrast	Est.	SE	90% CI	ECI	ECI/ σ_Y	TOST P ($0.20\sigma_Y$, BH)	$< 0.20\sigma_Y$
Passing Bogus Items	2-1	1.07 pp	0.84 pp	[-0.30, 2.45]	2.45 pp	0.058	< 0.001	Yes
	3-1	-0.57 pp	0.82 pp	[-1.93, 0.79]	1.93 pp	0.046	< 0.001	Yes
Passing IRIs	2-1	1.33 pp	0.84 pp	[-0.05, 2.72]	2.72 pp	0.060	< 0.001	Yes
	3-1	0.63 pp	0.85 pp	[-0.76, 2.02]	2.02 pp	0.044	< 0.001	Yes
Passing FMCs	2-1	0.25 pp	0.85 pp	[-1.16, 1.65]	1.65 pp	0.033	< 0.001	Yes
	3-1	0.12 pp	0.83 pp	[-1.25, 1.49]	1.49 pp	0.030	< 0.001	Yes
Exhibiting Straightlining	2-1	-0.38 pp	0.97 pp	[-1.98, 1.23]	1.98 pp	0.066	< 0.001	Yes
	3-1	0.41 pp	1.02 pp	[-1.27, 2.10]	2.10 pp	0.070	< 0.001	Yes
Answering Open-Ended Item	2-1	0.12 pp	0.42 pp	[-0.56, 0.80]	0.80 pp	0.052	< 0.001	Yes
	3-1	0.00 pp	0.38 pp	[-0.62, 0.62]	0.62 pp	0.040	< 0.001	Yes
Reading Time	2-1	-0.67 s	0.96 s	[-2.25, 0.92]	2.25 s	0.034	< 0.001	Yes
	3-1	-0.18 s	0.91 s	[-1.67, 1.32]	1.67 s	0.025	< 0.001	Yes
Treatment Effect	2-1	-0.05 pp	1.41 pp	[-2.37, 2.27]	2.37 pp	0.060	< 0.001	Yes
	3-1	-0.17 pp	1.35 pp	[-2.40, 2.05]	2.40 pp	0.061	< 0.001	Yes

Notes: Estimates are stratified pairwise differences in the marginal effect of a one-category increase in professionalism. The data-quality models estimate a separate professionalism slope in each observed item-position cell; the treatment model estimates the treatment-by-professionalism slope in each experiment-position cell. Contrasts are formed within strata observed at both positions and aggregated using sample-size weights. Binary outcomes and treatment effects are in percentage points; reading time is in seconds. TOST P values are BH-adjusted across the 14 contrasts in this table. No conventional contrast is statistically significant before or after BH adjustment (all unadjusted $P \geq 0.11$; all adjusted $P = 0.994$).

Supplementary Note 4: Sensitivity to Model-based Extrapolation and Cross-survey Weighting

Supplementary Table 7 compares three ways of forming contrasts identified in at least two surveys. The additive-model contrast is computed from the additive model used for the main figures. It adjusts predictions to a fixed pooled composition and uses the additive survey- and item- or experiment-specific position structure to predict cells not observed by design. The other two approaches restrict each contrast to survey-item or survey-experiment strata observed at both positions. The inverse-variance-weighted (IVW) approach first averages supported strata by observation count within survey and then combines survey estimates by inverse-variance weighting. The sample-size-weighted approach directly weights all supported strata by their representation in the pooled analytic sample while propagating respondent-clustered covariance within survey. The IVW and sample-size-weighted approaches therefore differ only in cross-survey weighting, whereas the additive-model approach also relies on model structure to bridge unsupported cells. Both stratified summaries are calculated separately for each position pair. Because the included strata and normalized weights can differ across pairs, they are not constrained to satisfy transitivity; a 3–1 contrast need not equal the sum of the corresponding 3–2 and 2–1 contrasts.

The three approaches lead to the same substantive conclusions. The IVW and sample-size-weighted estimate are nearly identical for most contrasts; their largest difference is 3.1 s for the Position-2-versus-1 reading-time contrast. Among percentage-point outcomes, the additive-model estimate differs from the sample-size-weighted estimate by at most 1.2 points, and the IVW estimate differs from it by at most 0.7 points.

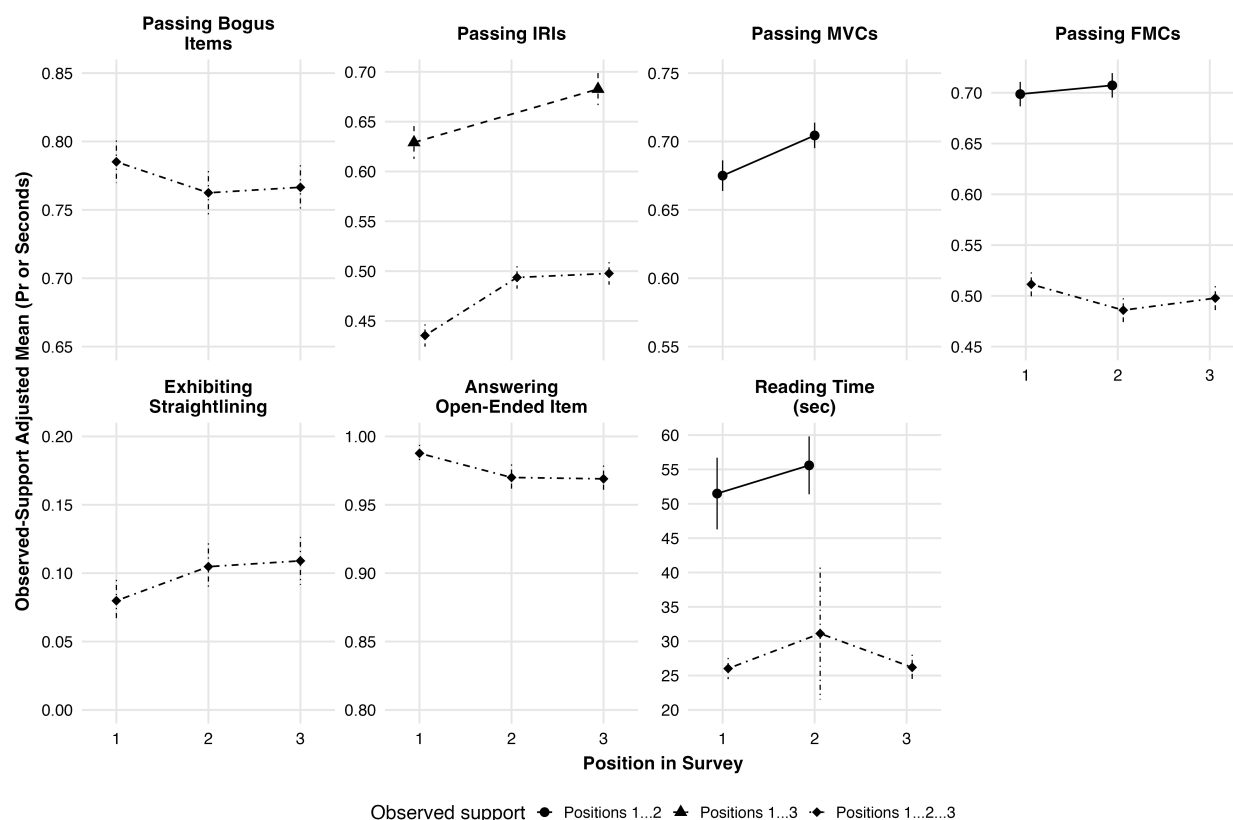
Supplementary Table 7: Comparison of additive-model, inverse-variance-weighted (IVW) and sample-size-weighted estimates.

Measure	Contrast	<i>k</i>	Surveys	Additive model (SE)	IVW Est. (SE)	Stratified (SE)	Add. -strat.	IVW -strat.
Passing FMCs	2-1	3	1, 3, 5	-1.41 (0.59)	-0.67 (0.54)	-1.41 (0.59)	0.00	0.74
	3-1	2	3, 5	-0.29 (0.80)	-1.37 (0.79)	-1.36 (0.79)	1.07	0.00
Passing IRIs	2-1	3	3, 4, 5	5.30 (0.88)	5.88 (0.71)	5.82 (0.72)	-0.52	0.06
	3-1	4	2, 3, 4, 5	5.96 (0.55)	5.98 (0.53)	5.96 (0.55)	0.00	0.03
Reading Time	2-1	3	1, 3, 5	4.63 (3.00)	1.50 (1.13)	4.61 (2.99)	0.02	-3.11
	3-1	2	3, 5	-1.96 (4.33)	0.49 (1.05)	0.16 (1.11)	-2.12	0.32
Treatment Effect (pooled)	2-1	3	1, 3, 5	-0.11 (0.98)	0.39 (0.94)	-0.10 (0.98)	-0.01	0.49
	3-1	2	3, 5	-1.18 (1.47)	-1.09 (1.30)	-1.10 (1.30)	-0.08	0.01
Treatment Effect: Diseases	2-1	2	3, 5	-2.09 (2.89)	-2.04 (2.88)	-2.04 (2.89)	-0.05	0.00
	3-1	2	3, 5	-1.73 (2.89)	-1.75 (2.89)	-1.72 (2.89)	-0.02	-0.03
Treatment Effect: Student Loans	2-1	3	1, 3, 5	-0.39 (1.72)	-0.25 (1.71)	-0.54 (1.73)	0.15	0.30
	3-1	2	3, 5	-0.20 (2.05)	-1.31 (2.11)	-1.40 (2.11)	1.20	0.08
Treatment Effect: Welfare	2-1	3	1, 3, 5	-0.15 (1.41)	0.02 (1.39)	-0.01 (1.41)	-0.14	0.03
	3-1	2	3, 5	-1.37 (1.72)	-0.18 (1.74)	-0.18 (1.74)	-1.19	-0.01

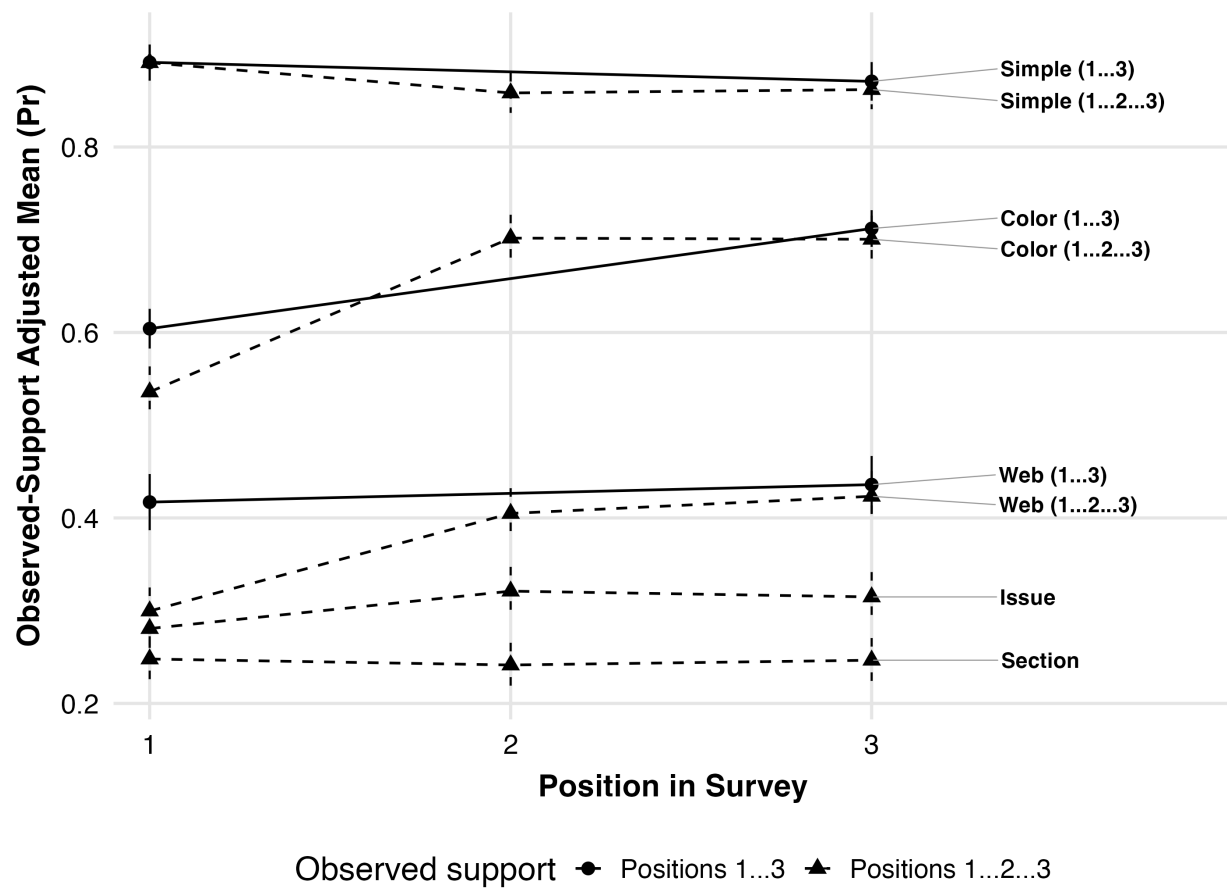
Notes: *k* is the number of surveys contributing supported strata; survey numbers correspond to Table 1. The additive-model column reports the contrast implied by the additive model used for the main figures. The IVW column uses the same supported strata but combines surveys by inverse-variance (fixed-effects meta-analytic) weighting. The stratified column uses sample-size weights and is the approach used for all formal inference in the main text. The IVW and stratified columns are comparison-specific and need not be transitive because the supported strata and normalized weights may differ across position pairs. Estimates and standard errors are in percentage points except for reading time, which is in seconds. Measures identified in only one survey are omitted because cross-survey aggregation is not required.

Supplementary Note 5: Observed-support Figures

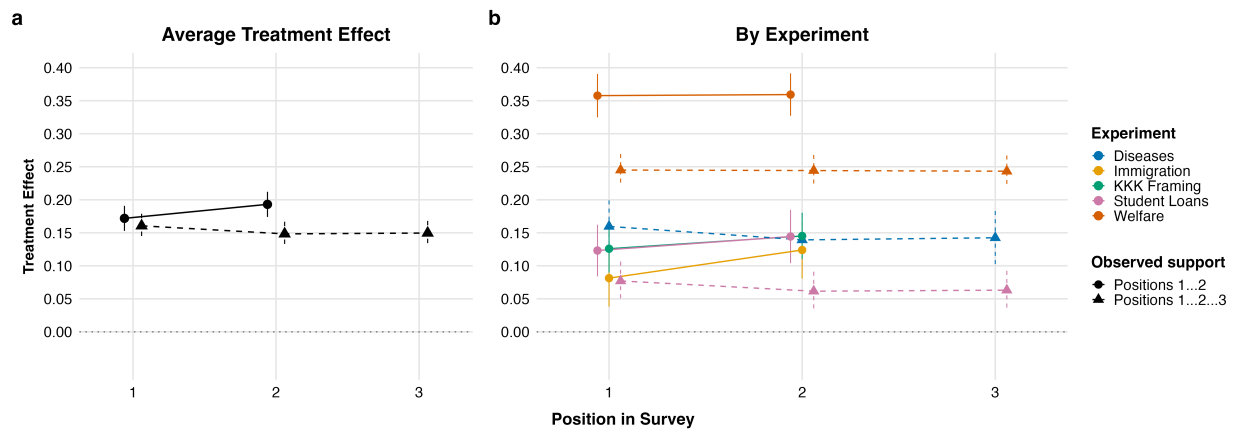
Supplementary Figures 1–4 present observed-support versions of the main-text figures: the adjusted means are recomputed using only the survey–item or survey–experiment strata that enter the corresponding pairwise stratified comparisons. The gap between each observed-support profile and its main-text counterpart shows where the plotted profiles depend on model-based predictions for cells not observed by design, and why the plotted profiles and the formal pairwise contrasts need not coincide exactly. The professionalism analyses are not repeated here since Surveys 4 and 5 fielded all three positions and Fig. 5 involves no model-based extrapolation.



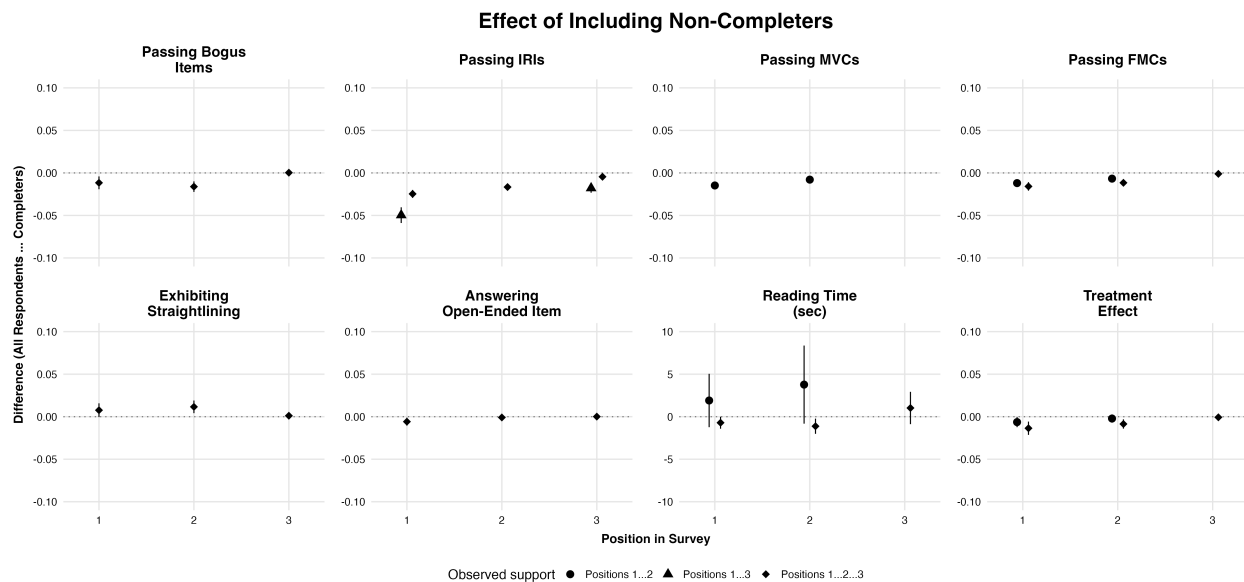
Supplementary Fig. 1: **Observed-support adjusted means for the data-quality outcomes.** Each line and marker type uses only the survey-item strata observed at the indicated positions (1-2, 1-3 or 1-2-3). Within each subset, means are adjusted to that subset's empirical composition and plotted only at the positions it identifies, so no point relies on model-based extrapolation. Error bars show 95% confidence intervals based on respondent-clustered standard errors.



Supplementary Fig. 2: **Observed-support IRI item profiles.** Adjusted pass rates for individual IRIs, shown separately by position-support set. For Color, Simple and Web, which were fielded in surveys with different position supports, separate lines are shown for the Positions 1-and-3 support (Survey 2) and the all-three-position support (Survey 3, plus Survey 4 for Color and Web). Items fielded in only one survey (Issue and Section) appear as a single line. Within each line, observed survey-item-position cells are averaged using fixed survey weights, so no point relies on model-based extrapolation. Error bars show 95% confidence intervals based on respondent-clustered standard errors.



Supplementary Fig. 3: **Observed-support adjusted treatment effects.** **a**, Pooled treatment effects by position for the two support sets that identify treatment-effect contrasts (positions 1–2 and 1–2–3). **b**, Experiment-specific profiles on the corresponding observed supports; points appear only at positions observed for that experiment. Outcomes are scaled 0–1 and oriented so that positive values correspond to the treatment direction in the original study. Error bars show 95% confidence intervals based on respondent-clustered standard errors.



Supplementary Fig. 4: **Observed-support attrition differences.** Differences between available-response and completer-only adjusted estimates at each position, shown separately by the sets of strata underlying the stratified comparisons; positive values indicate larger estimates when non-completers are included. The treatment-effect panel is pooled across experiments. Error bars show 95% confidence intervals based on respondent-clustered standard errors.