# Simulating Social Fingerprints: A DIF-Based Analysis of Structural Bias in LLM Responses

**Anonymous ACL submission**

## Abstract

Large language models (LLMs) are increasingly used to simulate survey respondents, yet their validity for social scientific inference remains underexplored. We apply psychometric methods—Item Response Theory (IRT) and Differential Item Functioning (DIF)—to assess whether LLM personas preserve the structural relationship between demographics and political attitudes. Analyzing 8,280 personas on ANES 2020 items, we find systematic distortions invisible to standard validation: roughly half of significant demographic bias effects reverse sign (e.g., gender favoring males instead of females), others amplify 2–5× or vanish, and LLM ideology distributions compress to one-fifth the human range. These patterns persist in our tested model, indicating that current LLMs fundamentally misrepresent subgroup attitude structures despite surface-level plausibility. Our anchor-based framework provides a rigorous diagnostic for human–AI alignment in survey contexts, revealing that LLM-generated responses are unsuitable for applications requiring faithful demographic inference—from hypothesis testing to fairness auditing—until these structural biases are resolved.

## 1 Introduction

Large language models (LLMs) are rapidly expanding their role as synthetic survey respondents for public opinion estimation, policy evaluation, and social science research (Argyle et al., 2023; Horton, 2023; Ziems et al., 2024). Their advantages are substantial: instant data collection, large-scale samples, precise demographic targeting, and the ability to simulate responses from hard-to-reach populations. Recent studies have shown that when conditioned on demographic profiles, LLMs can approximate aggregate human response distributions, suggesting potential as an effective alternative to traditional surveys (Argyle et al., 2023; Salecha et al., 2024).

However, current evaluations focus primarily on aggregate metrics—distributional similarity, group mean comparisons, and correlations with human averages (Bai et al., 2024; Hu et al., 2025). Such approaches can obscure critical *item-level* distortions. For instance, an LLM might correctly predict that "60% of conservatives support a given policy" while systematically misrepresenting *how* subgroups (by gender, race, education, etc.) *within* that ideological category form their responses. Ultimately, without examining group differences at the item level, measurement invariance cannot be guaranteed, potentially undermining the validity of group comparisons and opinion interpretation in systematic ways (Aera, 2014).

For example, two respondents might show similar views on government spending, healthcare, and environment—placing them at the same ideological position ($\theta$)—yet consistently diverge on gun control items, differing *only by gender*. To systematically detect such patterns, we employ Item Response Theory (IRT) (Rasch, 1993; Lord, 1952) and Differential Item Functioning (DIF) (Holland and Wainer, 2012). Specifically, we apply the Graded Response Model (GRM) (Samejima, 1969), an IRT model for ordinal responses. GRM estimates latent ideology $\theta$ from individual response patterns, placing everyone on a common scale. DIF then tests whether response probabilities for specific items differ across groups (gender, race, education, etc.) *at identical $\theta$ values*. Though standard in large-scale surveys like American National Election Studies (ANES) (The American National Election Studies, 2021), this framework has rarely been applied to compare humans and LLMs on equal footing.

This study addresses two questions:

1. Do LLMs reproduce human bias patterns on individual survey items?

2. Can IRT and DIF methods developed for hu-

mans detect such biases in LLMs?

We construct 8,280 demographically diverse personas based on the ANES 2020 dataset and simulate their responses to policy-related items using LLMs. Using IRT to estimate latent ideological traits and DIF to assess item-level bias, we find that LLMs exhibit bias directions similar to humans but with substantially amplified magnitudes (2–5×). Political ideology and religion emerge as the most influential attributes in both populations.

This study makes three contributions. First, it provides the first systematic application of psychometric methods (IRT and DIF) to LLM-simulated survey responses in a persona-based context. Second, it reveals that while LLMs reproduce human-like directional patterns of structural bias, they exhibit significantly amplified effect sizes, raising important questions about their validity as survey respondents. Third, it demonstrates that item-level psychometric analysis can provide interpretable, attribute-specific assessments of human-LLM alignment, offering a methodological foundation for future research in this domain.

## 2 Related Work

### 2.1 LLM Survey Simulation: Progress and Limitations

Human survey responses exhibit systematic biases—social desirability bias (favoring socially acceptable answers), acquiescence bias (tendency to agree), and framing effects (sensitivity to question wording) (Crowne and Marlowe, 1960; Winkler et al., 1982; Tversky and Kahneman, 1981). These biases complicate survey interpretation, motivating decades of psychometric research on detection and correction methods (Paulhus, 1991).

Recent work demonstrates that LLMs can approximate not only aggregate human response patterns but also some of these biases when conditioned on demographic profiles. Argyle et al. (2023) showed that GPT-3 accurately reproduces group-level distributions across U.S. electoral surveys, capturing how attitudes vary by ethnicity, gender, and age—suggesting that training data implicitly encodes diverse subgroup patterns. Similarly, Salecha et al. (2024) found that LLMs exhibit social desirability bias on Big Five personality assessments, mirroring human tendencies toward favorable self-presentation. These findings suggest that sufficiently advanced LLMs can replicate certain human response biases at the population level (Qu and Wang, 2024).

However, not all biases transfer uniformly. Tjuatja et al. (2024) tested whether LLMs replicate known framing effects from survey methodology and found that most models, especially those fine-tuned with RLHF, fail to reproduce human sensitivity to question wording. Instead, they show inconsistent or exaggerated reactions, suggesting that safety tuning may suppress certain biases. Bisbee et al. (2024) compared ChatGPT-generated survey responses to real human data and observed substantially reduced variance and skewed distributions, leading to different statistical properties in downstream analyses. These divergences may stem from training data composition—such as overrepresentation of Western, progressive viewpoints—or deliberate bias suppression during RLHF (Santurkar et al., 2023; Wang et al., 2025).

Critically, these studies share a methodological limitation: they evaluate LLMs using *aggregate metrics*—distributional similarity, group means, or correlations—that cannot isolate *item-level* distortions. If an LLM matches overall response rates but reverses *how* demographics shape individual items (e.g., gender effects on gun control flip from female-progressive to female-conservative), standard evaluations will miss the error. Without item-level measurement invariance (Millsap, 2012), we cannot determine whether LLMs preserve the structural relationships between demographics and attitudes—a prerequisite for valid subgroup comparisons. This gap motivates our psychometric approach.

### 2.2 Psychometric Foundations: IRT and DIF for Item-Level Analysis

Comparing survey responses across populations faces a fundamental challenge: **scale alignment**. If Person A (human) selects "5" and Person B (LLM) selects "3" on a 7-point ideology scale, does this mean B is more liberal, or does B simply use the scale differently—for instance, avoiding extreme options? Without a common measurement framework, we cannot distinguish genuine attitude differences from scale usage artifacts (Embretson and Reise, 2013).

Item Response Theory (IRT) solves this by estimating each respondent's *latent trait*—an unobservable characteristic like political ideology, denoted $\theta$—from their pattern of responses across multiple items (Lord, 1952; Rasch, 1993). For example, someone who consistently selects conservative op-

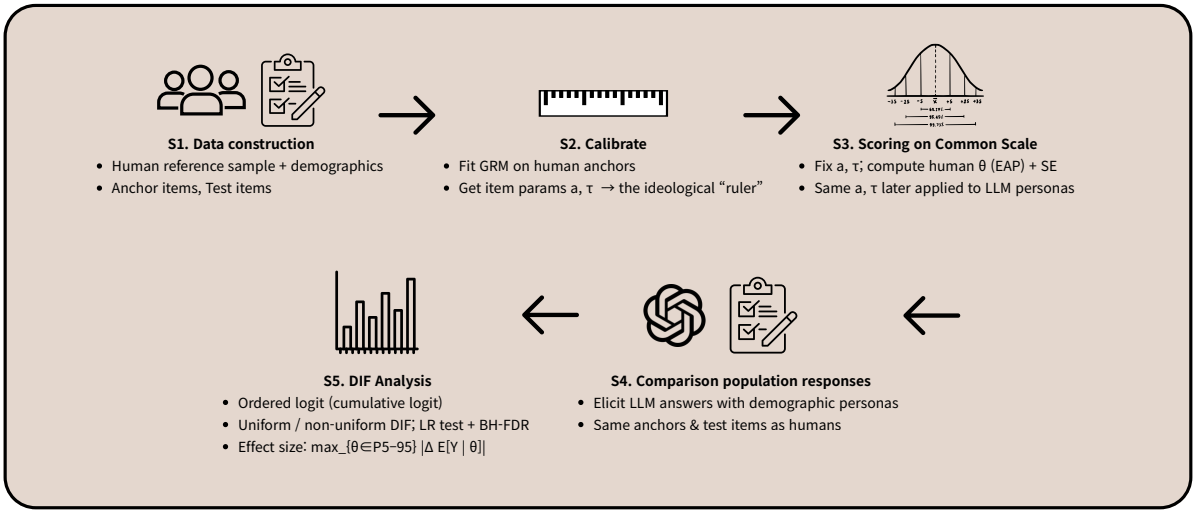


Figure 1: Overview of our five-stage, anchor-based pipeline: calibrate a shared ideology scale on humans, score the comparison population on that scale, then test item-level group differences (DIF).

tions receives a high $\theta$ score, regardless of whether they use "7" or "5" to express that position. This places all respondents on a common scale, analogous to converting Fahrenheit and Celsius to a single temperature metric.

IRT models characterize each survey item by two types of parameters (Baker and Kim, 2004): *discrimination* ($a$), which measures how strongly the item differentiates between respondents with different $\theta$ values, and *thresholds* ($\tau$), which specify the $\theta$ levels where respondents transition between response categories. A well-designed survey uses *anchor items*—questions with stable, group-neutral measurement properties—to establish this common scale (Brennan, 2014). Once anchors define the "ruler," we can score any new respondent (human or LLM) on the same $\theta$ axis and examine how they respond to *test items*—questions of substantive interest that may exhibit group differences.

With everyone on a common scale, Differential Item Functioning (DIF) tests whether demographic groups respond differently to items *after* controlling for $\theta$ (Holland and Wainer, 2012; Zumbo, 2007). Consider two respondents—one male, one female—who both score $\theta = 0.5$ (moderate conservative). If they systematically answer a gun control item differently, the item exhibits *gender DIF*. DIF can be *uniform* (a constant shift: one group always scores higher) or *non-uniform* (the gap changes across $\theta$: e.g., gender matters more among moderates than extremists) (Mellenbergh, 1989). Statistically, DIF is detected by testing whether adding group indicators improves prediction of item responses beyond $\theta$ alone (Crane et al., 2006).

These methods are standard in educational testing (Dorans et al., 2010) and survey research (The American National Election Studies, 2021), where they ensure that test scores remain comparable across demographic subgroups—a property called *measurement invariance* (Vandenberg and Lance, 2000). However, they have not been systematically applied to compare human and LLM populations. Existing LLM evaluation studies use aggregate metrics that conflate scale differences with genuine bias, making it impossible to isolate item-level distortions. We adapt the IRT/DIF framework to test whether LLMs preserve the structural relationships between demographics and attitudes at the item level—placing both populations on a shared ideological scale via anchor items, then using DIF to quantify where and how their response patterns diverge.

## 3 Methods

### 3.1 Overview

We apply a five-stage, anchor-based psychometric pipeline for comparing structural bias between two

populations. The key idea is to first place both populations on the *same latent ideology* scale, then test item-level group differences on that common scale. Without this alignment, apparent gaps may reflect measurement artifacts rather than genuine bias.

**Stage 1: Data construction.** Prepare (i) a *reference* population with demographic diversity, (ii) *anchor* items to define the common scale, and (iii) *test* items to probe group differences.

**Stage 2: Anchor calibration (GRM).** Fit a Graded Response Model(Samejima, 1969) to the reference population's anchor responses; the estimated item parameters define the ideological "ruler" used thereafter.

*Outputs:* fixed anchor parameters $(a, \tau)$.

**Stage 3: Scoring on the common scale (EAP).** With anchor parameters fixed, compute person scores ($\theta$) and standard errors for the reference population. The same parameters are later applied to the comparison population once responses are available.

**Stage 4: Comparison population responses.** Obtain responses from the comparison population—either by collecting data (e.g., another human cohort) or by simulating with personas (e.g., LLMs)— using the same anchor and test items.

**Stage 5: Item-level DIF analysis.** On the shared $\theta$ scale, test whether demographic groups respond differently to each *test item*. We use an ordinal logistic regression (cumulative logit model)(McCullagh, 2019), likelihood-ratio testing, FDR control(Benjamini and Hochberg, 1995), and an interpretable effect size: the *maximum absolute* difference in expected scores between groups over the central $\theta$ range (5th–95th percentiles).

This design is dataset- and model-agnostic; we instantiate it in Sections 3.2–3.6 (see Fig.1 for a schematic).

## 3.2 Data construction

### 3.2.1 ANES 2020 Dataset

We use the ANES 2020 Time Series (pre-election wave) (The American National Election Studies, 2021). The sample includes 8,280 respondents and covers key demographics (gender, race, education, income, age). ANES offers well-designed policy/value batteries that measure ideology reliably and can be mirrored for LLM personas, making it suitable for human–LLM comparison. Preprocessing is minimal: standardize text labels (drop numeric codes; keep age numeric) and map special/missing codes to NaN. As groupings for DIF we use standard binary splits (gender; race: White vs. non-White; education: BA+ vs. <BA; age bands); the full label-to-group mapping is in Appendix A.

### 3.2.2 Item selection

**Anchor items** define the common ideology scale shared by humans and LLMs. We select six 7-point items with minimal DIF in the human sample and use them to build the shared ruler (Sec. 3.3) and compute person scores on that scale (Sec. 3.4).

**Test items** assess whether groups (gender, race, education, age) differ in responses at the same ideology level. We choose eleven 2–5 point items on political/cultural issues, requiring (i) clear ordinal structure (ii) theoretical sensitivity to demographics, and (iii) no content overlap with anchors. All items are oriented so larger values indicate more conservative positions. Item texts (and response category counts) are summarized in Appendix B.

## 3.3 Anchor Calibration

### 3.3.1 GRM in brief

We use the Graded Response Model (GRM) (Samejima, 1969) to turn 7-point anchor items into a single latent ideology scale. For each item $i$ and category $m$, the model links higher ideology $\theta$ to higher response categories via

$$\Pr(Y_i \geq m \mid \theta) = \sigma\big(a_i(\theta - \tau_{i,m-1})\big), \quad (1)$$

$$\Pr(Y_i = m \mid \theta) = \Pr(Y_i \geq m \mid \theta) - \Pr(Y_i \geq m{+}1 \mid \theta). \quad (2)$$

Here $\sigma(\cdot)$ is the logistic function, $a_i$ controls how sharply responses shift, and $\tau_{i,m}$ are ordered category cutpoints.

### 3.3.2 Anchor Calibration

We fit the GRM on human responses to the anchor set to obtain item parameters $\{a_i, \tau_{i,\cdot}\}$ that define a shared "ruler." We then *fix* these parameters and score both humans and LLMs on the same latent scale so that the same score means the same ideology level in both populations. Anchors were pre-screened to be free of large group-specific bias (DIF); item directions were aligned so that larger values consistently indicate more conservative positions. For identification we center and scale the latent trait on the human reference population ($\theta \sim \mathcal{N}(0, 1)$) and align item polarity so

that larger categories indicate more conservative positions. Anchors were pre-screened on humans: any item with residual DIF effect size $> 0.5$ (after controlling for $\theta$) was excluded. Diagnostics for the final anchor set (information curves, PPC, residual DIF) are summarized later in Section 4.1(Baker and Kim, 2004; Embretson and Reise, 2013).

### 3.4 Theta Scoring (EAP)

We use the calibrated anchors from Section 3.3 as a shared ideological *ruler*. Holding the anchor item parameters fixed, we locate each respondent—human or LLM persona—on the same latent scale by *Expected A Posteriori (EAP)* scoring (Bock and Aitkin, 1981).

Let $\mathcal{A}$ be the anchor set and $\mathbf{y}_p = \{y_{pi}\}_{i\in\mathcal{A}}$ the anchor responses for person $p$. On a Gauss–Hermite grid with $G$=41 nodes, $\{(\theta_g, w_g)\}_{g=1}^{G}$ (Teukolsky et al., 1992), the (unnormalized) posterior weights are

$$\tilde{w}_{pg} \;\propto\; w_g\,\phi(\theta_g)\prod_{i\in\mathcal{A}}\Pr\big(Y_i{=}y_{pi}\mid\theta_g;\,\hat{a}_i,\hat{\tau}_{i,\cdot}\big), \tag{3}$$

where $\phi$ is the standard normal prior and the category probabilities follow the GRM in Sec. 3.3.1. For numerical stability we accumulate in log-space, $\log\tilde{w}_{pg} = \log w_g + \log\phi(\theta_g) + \sum_{i\in\mathcal{A}}\log\Pr(\cdot)$, shift by $\max_g \log\tilde{w}_{pg}$, exponentiate, and normalize to obtain $w_{pg} = \tilde{w}_{pg}/\sum_h \tilde{w}_{ph}$. The EAP score and its posterior standard error are then

$$\hat{\theta}_p^{\mathrm{EAP}} \;=\; \sum_{g=1}^{G}\theta_g\,w_{pg}, \tag{4}$$

$$\widehat{\mathrm{SE}}(\hat{\theta}_p)^2 \;=\; \sum_{g=1}^{G}(\theta_g-\hat{\theta}_p^{\mathrm{EAP}})^2\,w_{pg}. \tag{5}$$

Intuitively, the anchors define the ruler; Eq. (3) asks where on that ruler the observed anchor pattern is most plausible, and the EAP takes the plausibility-weighted average.

### 3.5 LLM Simulation

We elicit comparison-population responses by prompting LLM (gpt-4o-mini) with personas that mirror the reference labels. The *system* prompt frames the respondent as a U.S. adult with the given profile and requests a single numeric choice per item; the *user* prompt supplies the persona labels and the list of anchor and test items with their allowed options.

Per run and persona we run eight draws (temperatures 0.15 and 0.35, four seeds). Ties are broken by the lower-temperature outcome; unresolved ties are set to missing. The resulting response matrix mirrors the human item structure and is scored on the fixed-anchor scale (Secs. 3.3–3.4) before DIF analysis (Sec. 3.6).

### 3.6 DIF Model (Ordered Logit)

Given that Sections 3.3–3.4 place everyone on a common ideology axis $\theta$, we now ask for each *test item*: do two demographic groups respond differently *even at the same* $\theta$? We fit an ordered logit (cumulative logit) (McCullagh, 2019) with item thresholds $\tau_k$ and slopes, using a binary group indicator $g$ (majority vs. others):

$$\Pr(Y\le k\mid\theta,g) = \mathrm{logit}^{-1}\begin{pmatrix}\tau_k-\beta_\theta\,\theta-\beta_g\,g\\ -\,\beta_{\theta:g}\,\theta g\end{pmatrix}. \tag{6}$$

Uniform DIF corresponds to $\beta_g \neq 0$ (a constant shift), and non-uniform DIF to $\beta_{\theta:g}\neq 0$ (a gap that changes across ideology). We compare against the reduced model without group terms via a likelihood-ratio test and control FDR across items with Benjamini–Hochberg.

Per item we report: LR $p$ and BH-adjusted $q$, the presence/direction of uniform/non-uniform DIF, and an interpretable effect size: the *maximum absolute* difference in expected scores between groups over the central $\theta$ range (5th–95th percentiles). This single number summarizes how large the group gap can get on the response scale. Implementation notes: we check minimum counts for rare categories, separate thresholds from coefficients as in `statsmodels`, and guard against design-matrix name mismatches during prediction. These outputs feed into topic- and group-level summaries and human–LLM contrasts on the same $\theta$ scale. Unless otherwise noted, we declare significance at BH-adjusted $q{<}0.01$ and report whether uniform ($\beta_g\neq 0$) and/or non-uniform ($\beta_{\theta:g}\neq 0$) components are present.

Anonymized codes are provided for reproduction (see Appendix C: Code Availability).

## 4 Results

### 4.1 Anchor Validation

We first confirm that the six anchors act as a reliable common ruler. Figure 2 plots cumulative test

| Item | $a$ | $[\tau_{\min}, \tau_{\max}]$ | PPC↓ | rDIF | $\theta^*$ |
|---|---|---|---|---|---|
| V201200 | 3.33 | $[-1.45, 1.40]$ | 71.25 | 0.46 | $-0.45$ |
| V201246 | 2.92 | $[-0.89, 1.37]$ | 45.09 | 0.42 | 0.85 |
| V201252 | 3.30 | $[-0.71, 0.90]$ | 36.41 | 0.24 | $-0.20$ |
| V201255 | 3.48 | $[-1.05, 0.87]$ | 42.01 | 0.37 | 0.35 |
| V201258 | 3.06 | $[-0.85, 0.97]$ | 47.96 | 0.41 | 0.55 |
| V201262 | 3.31 | $[-0.45, 1.24]$ | 41.71 | 0.26 | 0.05 |

Table 1: Anchor diagnostics from the human GRM fit. $a$: discrimination. $[\tau_{\min}, \tau_{\max}]$: threshold range (coverage along the ideology axis). PPC: posterior predictive check (smaller is better). rDIF: residual DIF (max; all $< 0.5$). $\theta^*$: information peak location.

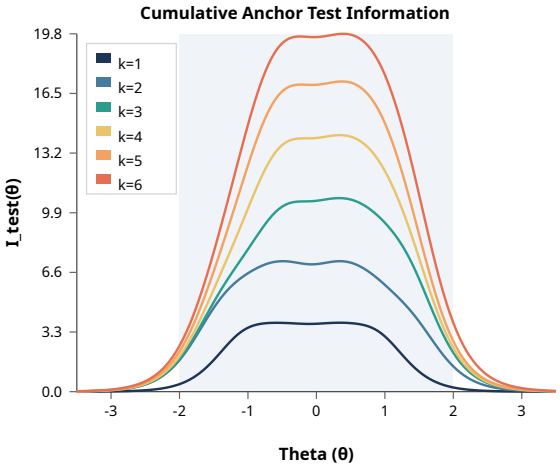


Figure 2: Cumulative test information $I_{\text{test}}(\theta)$ as anchors are added (sorted by discrimination $a$). Shaded core range: $\theta \in [-2, 2]$.

information as anchors are added in descending discrimination $a$. The curve shows adequate precision across the core range $\theta \in [-2, 2]$, which contains the bulk of respondents. This supports using the calibrated anchors to place both humans and LLM personas on the same latent ideology axis.

Table 1 summarizes per-item diagnostics. Discriminations are high ($a \in [2.92, 3.48]$); threshold spans cover the ideology continuum (e.g., $\tau_{\min}$ to $\tau_{\max}$ lie within $[-1.45, 1.40]$ across items); posterior predictive checks are acceptable (PPC $= 36.4$–$71.3$, with V201200 highest yet within tolerance); and residual DIF magnitudes are below the screening threshold ($< 0.5$). Information peaks fall in $[-0.45, 0.85]$, matching the high-information band in Fig. 2. Taken together, the anchors behave as a DIF-robust ruler for cross-population comparisons and motivate downstream DIF analysis on the shared $\theta$ scale.

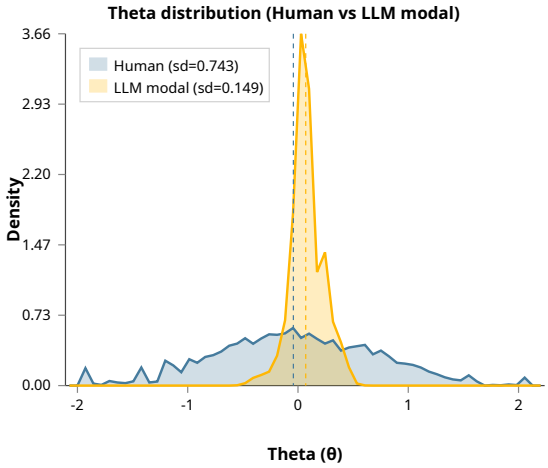


Figure 3: EAP $\theta$ distributions for Humans vs. LLMs on the same human-calibrated anchor scale. LLMs are sharply concentrated near the center ($\text{sd} = 0.149$, range $\approx [-0.57, 0.59]$) compared to Humans ($\text{sd} = 0.743$, range $\approx [-1.91, 2.04]$), with a slight conservative shift in the mean (Human $-0.032$ vs. LLM $+0.081$). Because anchor parameters are fixed from the human calibration, this compression reflects response behavior, not a scaling artifact.

## 4.2 Bias Magnitude and Direction

We compare human–LLM DIF at the item level on the shared $\theta$ scale. Figure 4 summarizes alignment using *signed expected-score gaps* on the observed response scale (reference group minus others; higher = more conservative). Points on the 45-degree line indicate agreement in both *direction* and *magnitude*, whereas Quadrants II/IV indicate *sign reversals*. Among pairs significant in *both* populations (BH–FDR $q<0.01$), we observe a mixed pattern: roughly half align in sign and half reverse. For example, education effects on `V201409` and `V201639` align (Humans: $+0.209/+0.640$; LLMs: $+0.175/+0.209$), while gender on `V201639` ($-0.324$ vs. $+0.521$) and race on `V201427` ($-0.378$ vs. $+0.464$) flip sign. We also see attenuation-to-zero for strong human effects (e.g., education on `V201640`, race on `V201640`), and several *amplified* LLM effects (e.g., gender on `V201409`: $-0.072$ vs. $-0.295$). Because these effects are on the *response* scale (not on $\theta$), discrepancies are not due to the compressed LLM $\theta$ variance but to different item-level mappings on the common ruler.

Table 2 complements the scatter by listing representative item–attribute pairs, reporting the human effect, the LLM effect on the same scale, a

| Item | Attribute (ref) | Human | LLM | Align | Amplif. |
|---|---|---|---|---|---|
| V201640 | Gender (Female) | $-0.267$ | $-0.122$ | Yes | $0.46\times$ |
| V201639 | Gender (Female) | $-0.324$ | $+0.521$ | No | $1.61\times$ (flip) |
| V201409 | Gender (Female) | $-0.184$ | $-0.072$ | Yes | $0.39\times$ |
| V201427 | Gender (Female) | $+0.060$ (n.s.) | $-0.216$ | n/a | — |
| V201639 | Race (White) | $-0.287$ | $-0.336$ | Yes | $1.17\times$ |
| V201409 | Race (White) | $-0.230$ | $+0.177$ | No | $0.77\times$ (flip) |
| V201427 | Race (White) | $-0.378$ | $+0.464$ | No | $1.23\times$ (flip) |
| V201409 | Education (<BA) | $+0.209$ | $+0.175$ | Yes | $0.84\times$ |
| V201639 | Education (<BA) | $+0.640$ | $+0.209$ | Yes | $0.33\times$ |
| V201427 | Education (<BA) | $+0.341$ | $+0.095$ | Yes | $0.28\times$ |
| V201639 | Age (35–59) | $-0.090$ | $+0.132$ | No | $1.47\times$ (flip) |
| V201409 | Age (35–59) | $+0.139$ | $-0.086$ | No | $0.62\times$ (flip) |

Table 2: Representative Human–LLM DIF on shared groupings (gender; race: White vs. non-White; education; age). Values are signed expected-score gaps (ref. group − others). "Amplif." is $|\Delta_{\text{LLM}}|/|\Delta_{\text{Human}}|$; "(flip)" marks sign reversal.

direction-alignment flag, and an amplification factor $|\Delta_{\text{LLM}}|/|\Delta_{\text{Human}}|$ ("flip" if signs differ).

### 4.3 Theta Distribution Divergence

Figure 3 compares EAP $\theta$ distributions for Humans and LLMs scored on the *same* human-calibrated anchors (Secs. 3.3–3.4). Two facts are immediate. First, LLM scores are sharply compressed around the center: $\text{sd}_{\text{LLM}}$=0.149 versus $\text{sd}_{\text{Human}}$=0.743 (about $5\times$ narrower), with much tighter support (LLM $\approx [-0.57, 0.59]$ vs. Human $\approx [-1.91, 2.04]$). Second, the LLM center is slightly shifted to the conservative side (means/medians: Human $-0.032/-0.042$; LLM $+0.081/+0.070$).

Because anchor parameters are fixed from the human calibration, identical $\theta$ has the *same substantive meaning* in both populations; the compression therefore reflects response behavior, not a scaling artifact. Practically, most LLM comparisons in Sec. 4.2 occur near $\theta \approx 0$, which can (i) attenuate non-uniform DIF that materializes chiefly in ideological tails, and (ii) over-represent central tendencies. Our effect-size—maximum absolute expected-score gap over the empirical 5th–95th $\theta$ percentiles—mitigates extrapolation to regions without data mass, but we caution that LLMs contribute limited information in the extremes of the ideology continuum.

## 5 Discussion

Our anchor-based DIF analysis reveals that LLMs encode demographic bias patterns that systematically diverge from human populations—not merely in magnitude, but in fundamental direction. Sign reversals occur in roughly half of significant item–attribute pairs: where human data show females responding more liberally (V201639: $-0.324$), LLM personas flip to favor males ($+0.521$), and similar inversions emerge for race effects (Table 2). This is not quantitative noise but qualitative misalignment, suggesting that persona prompts activate stereotypical associations that contradict empirical patterns (Argyle et al., 2023; Santurkar et al., 2023). Combined with the drastically compressed LLM ideology distribution ($\text{sd}_{\text{LLM}} = 0.149$ vs. $\text{sd}_{\text{Human}} = 0.743$)—likely driven by training objectives favoring consensus over viewpoint diversity (Perez et al., 2022; Bai et al., 2022)—these findings have immediate scholarly and practical implications.

**For researchers:** LLM-generated personas cannot serve as plug-in replacements for human respondents in survey simulation or hypothesis testing (Argyle et al., 2023). The observed bias inversions risk propagating erroneous assumptions about subgroup opinions, while the narrow $\theta$ range limits power to detect non-uniform DIF or stress-test instruments at ideological extremes. **For practitioners:** Deploying LLMs in fairness audits (Gupta et al., 2024) or synthetic data augmentation (Cheng et al., 2023) without accounting for these directional biases may inadvertently amplify rather than mitigate disparities. Future work should investigate whether enriched persona specifications, targeted debiasing (Gupta et al., 2024), or alternative prompting frameworks can realign LLM outputs with empirical demographic realities—and whether the ideological compression reflects an intrinsic limitation of current alignment methods or a reme-

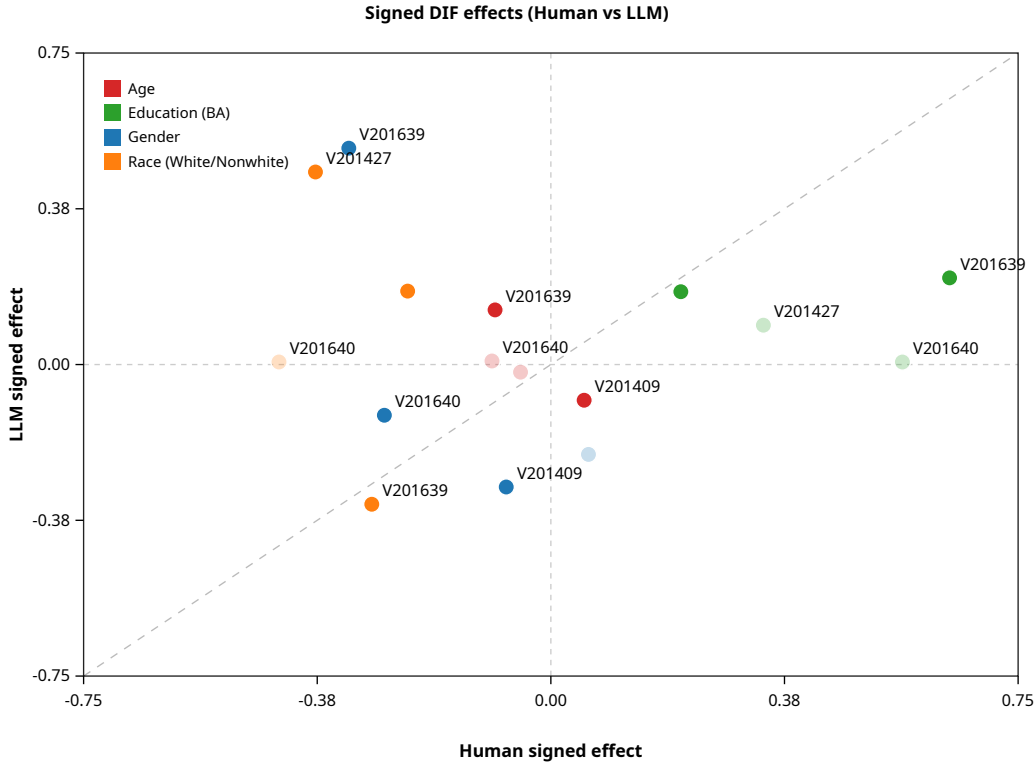


Figure 4: Item–attribute bias alignment between Humans (x-axis) and LLMs (y-axis). Each point is a test item $\times$ group (color-coded by attribute: Education, Race (White/Nonwhite), Gender, Age). Coordinates are signed expected-score gaps on the response scale (ref. group minus others; higher = more conservative), computed as in Sec. 3.6. Filled markers denote pairs significant for *both* Humans and LLMs (BH–FDR $q < 0.01$); half-filled markers indicate significance in exactly one population; hollow markers indicate non-significance in both. The 45-degree line marks perfect alignment. Quadrants II/IV indicate *sign reversals*.

diable artifact of RLHF design.

## 6 Conclusion

We apply psychometric methods (IRT/DIF) to expose a critical flaw in LLM survey simulations: personas systematically misrepresent demographic bias—not just in magnitude, but in direction. Across 8,280 personas on ANES 2020 items, we find that roughly half of significant bias effects flip sign relative to human data (e.g., gender favoring males instead of females), while others amplify 2–5$\times$ or vanish entirely. Compounding this, LLM ideology distributions compress to one-fifth the human range, clustering near the center and obscuring polarization. These distortions remain invisible to surface-level validation and threaten the validity of using LLMs for subgroup attitude inference or survey pretesting. Our anchor-based framework provides a principled diagnostic for human–AI alignment at the item level. The findings suggest that current LLMs, despite fluent outputs, fundamentally lack the structural relationship between demographics and attitudes required for credible social science applications. Future work must address whether this stems from training data, alignment processes, or architectural limits—and whether targeted interventions can restore demographic fidelity without sacrificing LLMs' efficiency gains.

## 7 Limitations

Our study has several important boundaries. **Domain and data:** We analyze U.S. political ideology (ANES 2020) using Likert-scale items; bias patterns may differ across cultures (e.g., East Asian collectivism vs. Western individualism), domains (consumer preferences, health attitudes), or question formats (open-ended, forced-choice). The six anchor items, while validated here, represent a small ideological slice and may not generalize to economic or social policy dimensions. **Models and methods:** We tested LLM (gpt-4o-mini) with standard demographic persona prompting—newer

models, chain-of-thought, or role-playing prompts might reduce bias. Our IRT framework assumes unidimensional latent traits and parametric response functions (GRM), which may oversimplify multidimensional attitudes or non-monotonic item behaviors. **Human baseline:** We benchmark against ANES human responses, which themselves contain measurement error, sampling bias (e.g., non-response), and social desirability effects; LLM biases are thus relative, not absolute. **Causality:** We document bias patterns but cannot isolate whether they originate from pretraining corpora, RLHF, persona interpretation, or prompt sensitivity. Despite these constraints, the consistent pattern in our tested model indicates a robust phenomenon warranting caution in deploying LLMs for sensitive social scientific inference until these issues are resolved.

## References


APA Aera. 2014. Ncme. *Standards for educational and psychological testing. Washington, DC: American Educational Research Association*, 1.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Xuechen Bai, Alex Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *Preprint*, arXiv:arXiv:2402.04105.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Frank B Baker and Seock-Ho Kim. 2004. *Item response theory: Parameter estimation techniques*. CRC press.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4):401–416.

R Darrell Bock and Murray Aitkin. 1981. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Robert L Brennan. 2014. *Test equating, Scaling, and Linking-Methods and practices*. Springer-verlag New York Incorporated.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.

Paul K Crane, Laura E Gibbons, Lance Jolley, and Gerald van Belle. 2006. Differential item functioning analysis with ordinal logistic regression techniques: Difdetect and difwithpar. *Medical care*, 44(11):S115–S123.

Douglas P Crowne and David Marlowe. 1960. A new scale of social desirability independent of psychopathology. *Journal of consulting psychology*, 24(4):349.

Neil J Dorans, Tim P Moses, and Daniel R Eignor. 2010. Principles and practices of test score equating. *ETS Research Report Series*, 2010(2):i–41.

Susan E Embretson and Steven P Reise. 2013. *Item response theory for psychologists*. Psychology Press.

S Gupta, V Shrivastava, A Deshpande, A Kalyan, P Clark, A Sabharwal, and T Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned llms (no. arxiv: 2311.04892). arxiv.

Paul W Holland and Howard Wainer. 2012. *Differential item functioning*. Routledge.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Tongtong Hu, Yevheniia Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Frederic Lord. 1952. A theory of test scores. *Psychometric monographs*.

Peter McCullagh. 2019. *Generalized linear models*. Routledge.

Gideon J Mellenbergh. 1989. Item bias and item response theory. *International journal of educational research*, 13(2):127–143.

Roger E Millsap. 2012. *Statistical approaches to measurement invariance*. Routledge.

Delroy L Paulhus. 1991. Measurement and control of response bias.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Georg Rasch. 1993. *Probabilistic models for some intelligence and attainment tests.* ERIC.

Aakanksha Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. 2024. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12):pgae533.

Fumiko Samejima. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1):1–97.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Saul A Teukolsky, Brian P Flannery, W Press, and W Vetterling. 1992. Numerical recipes in c. *SMR*, 693(1):59–70.

The American National Election Studies. 2021. The anes 2020 time series study. https://electionstudies.org/data-center/2020-time-series-study/. Accessed May 19, 2025.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? A case study in survey design. *arXiv preprint arXiv:2311.04076*.

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

Robert J Vandenberg and Charles E Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, 3(1):4–70.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

John D Winkler, David E Kanouse, and John E Ware. 1982. Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, 67(5):555.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Bruno D Zumbo. 2007. Three generations of dif analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2):223–233.

## A ANES Demographic Details

**Scope** This appendix reports the demographic composition of the ANES 2020 reference sample used for anchor calibration and DIF analysis.

**Tables** Raw-category distributions are shown in Tables 3–5 (Age/Education/Gender; Income/Marital; Race/Religion). Analysis-ready grouped distributions appear in Table 6.

**Notes** Non-substantive responses (e.g., *Don't know*, *Refused*) are excluded where applicable. Proportions are within-category shares and may not sum to 100% due to rounding. Grouping rules (e.g., BA+, age bands, race White vs. non-White) match those used in the DIF models.

**Packages (table rendering)** Main text / full-width one-page tables (Option B): `booktabs`, `array`, `ragged2e`, `threeparttable`, `adjustbox`. If a multi-page version is preferred (Appendix only), use: `booktabs`, `array`, `ragged2e`, `longtable`, `threeparttablex`. Optional numeric formatting: `siunitx`.

## B ANES Items Details

**Scope** This appendix lists the anchor and test items drawn from ANES 2020. Anchor items define the shared latent ideology scale; test items are used for DIF analysis across demographic groups.

**Table** See Table 7 for item codes, types (Anchor/Test), topical labels, summary wording, and the number of response categories.

**Notes** All items are oriented so that larger numeric responses indicate more conservative positions where required; polarity is aligned before estimation. Non-substantive responses are excluded from descriptive summaries and modeling.

## C Code Availability

An anonymous implementation of the simulation code and data processing scripts is available at the following URL:

https://anonymous.4open.science/r/project-anon-C033/

This repository has been anonymized in accordance with the ACL reviewing policy and will remain accessible during the review process.

Table 3: ANES 2020 Demographics — Raw (A: Age, Education, Gender)

| attribute | value | count | proportion |
|---|---|---|---|
| **Age** | | | |
| age_binned | 60s | 1562 | 0.19 |
| age_binned | 70+ | 1452 | 0.18 |
| age_binned | 30s | 1381 | 0.17 |
| age_binned | 50s | 1335 | 0.16 |
| age_binned | 40s | 1199 | 0.14 |
| age_binned | Under 30 | 1003 | 0.12 |
| age_binned | Missing | 348 | 0.04 |
| **Education** | | | |
| education | Some post-high school | 2790 | 0.34 |
| education | Bachelor's degree | 2055 | 0.25 |
| education | Graduate degree | 1592 | 0.19 |
| education | High school | 1336 | 0.16 |
| education | Less than high school | 376 | 0.05 |
| education | Missing | 131 | 0.02 |
| **Gender** | | | |
| gender | Female | 4450 | 0.54 |
| gender | Male | 3763 | 0.45 |
| gender | Missing | 67 | 0.01 |

Proportions may not sum to 100% due to rounding.

Table 4: ANES 2020 Demographics — Raw (B: Household Income, Marital Status)

| attribute | value | count | proportion |
|---|---|---|---|
| **Household income** | | | |
| household_income | Under $9,999 | 719 | 0.09 |
| household_income | Missing | 616 | 0.07 |
| household_income | $50,000–59,999 | 546 | 0.07 |
| household_income | $100,000–109,999 | 506 | 0.06 |
| household_income | $250,000 or more | 449 | 0.05 |
| household_income | $80,000–89,999 | 426 | 0.05 |
| household_income | $175,000–249,999 | 416 | 0.05 |
| household_income | $150,000–174,999 | 404 | 0.05 |
| household_income | $125,000–149,999 | 347 | 0.04 |
| household_income | $110,000–124,999 | 343 | 0.04 |
| household_income | $30,000–34,999 | 327 | 0.04 |
| household_income | $20,000–24,999 | 325 | 0.04 |
| household_income | $60,000–64,999 | 325 | 0.04 |
| household_income | $40,000–44,999 | 321 | 0.04 |
| household_income | $90,000–99,999 | 304 | 0.04 |
| household_income | $10,000–14,999 | 282 | 0.03 |
| household_income | $70,000–74,999 | 264 | 0.03 |
| household_income | $25,000–29,999 | 263 | 0.03 |
| household_income | $35,000–39,999 | 242 | 0.03 |
| household_income | $75,000–79,999 | 233 | 0.03 |
| household_income | $45,000–49,999 | 230 | 0.03 |
| household_income | $15,000–19,999 | 210 | 0.03 |
| household_income | $65,000–69,999 | 182 | 0.02 |
| **Marital status** | | | |
| marital_status | Married: spouse present | 4315 | 0.52 |
| marital_status | Never married | 1951 | 0.24 |
| marital_status | Divorced | 1221 | 0.15 |
| marital_status | Widowed | 567 | 0.07 |
| marital_status | Separated | 163 | 0.02 |
| marital_status | Missing | 56 | 0.01 |
| marital_status | Married: spouse absent | 7 | 0.00 |

Proportions may not sum to 100% due to rounding.

Table 5: ANES 2020 Demographics — Raw (C: Race, Religion)

| attribute | value | count | proportion |
|---|---|---|---|
| **Race** | | | |
| race | White | 5963 | 0.72 |
| race | Hispanic | 762 | 0.09 |
| race | Black | 726 | 0.09 |
| race | Asian | 284 | 0.03 |
| race | Multiple races, non-Hispanic | 271 | 0.03 |
| race | Native American | 172 | 0.02 |
| race | Missing | 102 | 0.01 |
| **Religion** | | | |
| religion | Protestant | 2113 | 0.26 |
| religion | Roman Catholic | 1640 | 0.20 |
| religion | Something else | 1555 | 0.19 |
| religion | Nothing in particular | 1475 | 0.18 |
| religion | Agnostic | 456 | 0.06 |
| religion | Atheist | 340 | 0.04 |
| religion | Jewish | 188 | 0.02 |
| religion | Orthodox Christian | 154 | 0.02 |
| religion | Latter-Day Saints | 113 | 0.01 |
| religion | Missing | 83 | 0.01 |
| religion | Buddhist | 74 | 0.01 |
| religion | Muslim | 48 | 0.01 |
| religion | Hindu | 41 | 0.00 |

Proportions may not sum to 100% due to rounding.

Table 6: ANES 2020 Demographics — Grouped for Analysis

| attribute | value | count | proportion |
|---|---|---|---|
| age_group | 35~59 | 3282 | 0.40 |
| age_group | Over 60 | 3014 | 0.36 |
| age_group | Under 35 | 1636 | 0.20 |
| age_group | Missing | 348 | 0.04 |
| education_ba_plus_group | <BA | 4502 | 0.54 |
| education_ba_plus_group | BA+ | 3647 | 0.44 |
| education_ba_plus_group | Missing | 131 | 0.02 |
| education_group | High | 3647 | 0.44 |
| education_group | Mid | 2790 | 0.34 |
| education_group | Low | 1712 | 0.21 |
| education_group | Missing | 131 | 0.02 |
| gender_group | Female | 4450 | 0.54 |
| gender_group | Male | 3763 | 0.45 |
| gender_group | Missing | 67 | 0.01 |
| income_group | Middle | 3129 | 0.38 |
| income_group | Low | 2919 | 0.35 |
| income_group | High | 1616 | 0.20 |
| income_group | Missing | 616 | 0.07 |
| marital_group | Married | 4315 | 0.52 |
| marital_group | Previously married | 1958 | 0.24 |
| marital_group | Never married | 1951 | 0.24 |
| marital_group | Missing | 56 | 0.01 |
| race_white_nonwhite_group | white | 5963 | 0.72 |
| race_white_nonwhite_group | nonwhite | 2215 | 0.27 |
| race_white_nonwhite_group | Missing | 102 | 0.01 |
| religion_group | Western religion | 4020 | 0.49 |
| religion_group | Non-religious | 2271 | 0.27 |
| religion_group | Other religion | 1555 | 0.19 |
| religion_group | Non-Western religion | 351 | 0.04 |
| religion_group | Missing | 83 | 0.01 |

Proportions may not sum to 100% due to rounding.

Table 7: ANES 2020 Anchor/Test Items

| Code | Type | Topic | Question (summary) | # Categories |
|---|---|---|---|---|
| V201200 | Anchor | Self-placed ideology | Where do you place your political ideology on a 1 (very liberal)–7 (very conservative) scale? | 7 |
| V201246 | Anchor | Government services vs. taxes | Which is closer to your view: more government services (higher taxes) vs. fewer services (lower taxes)? (1–7 continuum) | 7 |
| V201252 | Anchor | Health insurance (public vs. private) | Which is closer to your view: a government insurance plan vs. a private insurance plan? (1–7 continuum) | 7 |
| V201255 | Anchor | Jobs/welfare (government vs. individual) | Which is closer to your view: government should see to jobs and standard of living vs. people should get ahead on their own? (1–7 continuum) | 7 |
| V201258 | Anchor | Racial policy (help vs. self-reliance) | Which is closer to your view: the government should help Blacks vs. Blacks should help themselves? (1–7 continuum) | 7 |
| V201262 | Anchor | Environmental regulation (stronger vs. burden) | Which is closer to your view: tougher regulations on business are needed to protect the environment vs. regulations already place too much burden on business? (1–7 continuum) | 7 |
| V201639 | Test | Perceptions of sexism (innocuous remarks) | Degree of agreement with "Many women interpret innocent remarks or acts as sexist." (strongly agree–strongly disagree) | 5 |
| V201640 | Test | Gender roles / women's power | Degree of agreement with "Women try to gain power by controlling men." | 5 |
| V201403 | Test | Paid parental leave | Favor, oppose, or neither favor nor oppose requiring employers to provide paid leave to parents of newborns | 3 |
| V201409 | Test | Transgender bathroom policy | Should transgender people use bathrooms matching sex at birth, or bathrooms matching their self-identified gender? | 2 |
| V201406 | Test | Wedding services for same-sex couples | Allow refusal of services on religious grounds vs. require provision to same-sex couples | 2 |
| V201412 | Test | Employment anti-discrimination (LGBTQ) | Favor or oppose laws prohibiting employment discrimination against gays and lesbians | 2 |
| V201415 | Test | Adoption by same-sex couples | Should gay or lesbian couples be legally permitted to adopt children? (yes/no) | 2 |
| V201360 | Test | Felon voting rights (original) | Favor, oppose, or neither favor nor oppose allowing people with felony convictions to vote after completing their sentences | 3 |
| V201362x | Test | Felon voting rights (strength) | Strength of support for felon voting rights: 1 (strongly favor)–7 (strongly oppose) | 7 |
| V201418 | Test | End birthright citizenship | Favor, oppose, or neither favor nor oppose ending automatic birthright citizenship for children of undocumented immigrants | 3 |
| V201427 | Test | Importance of learning English | How important is it that everyone in the U.S. learns English? (very–somewhat–not very–not at all) | 4 |
| V201343 | Test | Death penalty | Favor or oppose the death penalty for those convicted of murder | 2 |

*Notes.* (i) Response counts (# Categories) exclude non-substantive codes (e.g., $-8$ Don't know, $-9$ Refused). (ii) Several 7-point anchors differ in ideological polarity across endpoints; analyses recode items so that higher values align with a consistent polarity when required.