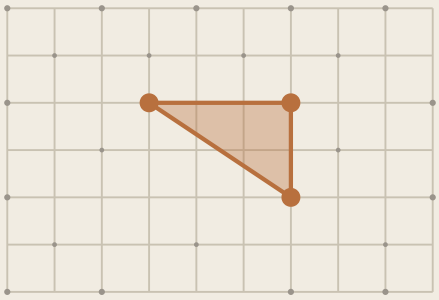
METHODOLOGY · PUBLIC EVAL

Replism against PEW ATP 163.

We ran our synthetic audience against the Pew Research Center's American Trends Panel and compared marginals, option by option, against published ground truth. *This document is the full result. The good, the bad, and what we're doing about both.*



ATP W163 · n=100 · weighted by WEIGHT_W163

WHY WE PUBLISH THIS

Synthetic audiences are easy to demo and hard to defend. We publish our evaluation work because *"trust the model"* is not a methodology. The numbers below are what the instrument did. Where it underperformed, we say so, and we say what we're doing about it.

WAVE	ATP W163
SOURCE	Pew Research Center
FIELD DATES	Feb 10 – 17, 2025
MODE	Online · Self-administered
FRAME	U.S. adults · 18+
REPLISM RUN	2026-05-14 · v0.1
WEIGHTING	WEIGHT_W163 · n=100
EXCLUSIONS	"Refused" removed

ACCURACY
93.6%
Against human self-replication.

MEAN OVERLAP
0.852
Population-level distribution overlap.

ITEMS > 0.90
21%
Replicating Pew to within a few points.

MEAN MAE/OPTION
7.9pp
Average per-option drift from ground truth, in percentage points.

What we mean by human self-replication. If you ask the same person the same question twice, they don't always give the same answer. Re-survey the same panel and the average overlap between the two waves is **91%**. That is the ceiling any instrument, human or synthetic, can plausibly hit. Replism's 93.6% reads as a share of that ceiling. https://replism.com/knowledgebase/human-self-replication

ABOUT THE DATASET

PEW ATP Wave 163: what it is and why we chose it.

The **American Trends Panel** is Pew Research Center's nationally representative, probability-based panel of U.S. adults. **Wave 163** is a multi-topic instrument fielded online to `n=5,089` respondents, covering social media use, race and racial issues, and views on gender. It is, in survey-research terms, a known quantity: published microdata, documented weighting, public crosstabs.

We picked it because it is **hard on purpose** and also the latest ATP wave published to date. The wave leans heavily on five-point intensity scales and on questions about race, territory where synthetic audiences have historically struggled with social desirability and with the tails of a distribution. If Replism can replicate ATP's marginals here, it earns the right to be trusted on novel research questions.

OVERLAP DISTRIBUTION — median 0.868 · σ 0.060 · range 0.737–0.982

≥ 0.90	6	20.7%
0.80 – 0.90	16	55.2%
0.70 – 0.80	7	24.1%
< 0.70	0	0.0%

Grounded.

Every Replism persona is anchored to real-world attitudinal microdata. No invented personalities.

Evaluated.

Each wave is scored item by item and the score is published. No private dashboards.

Honest about misses.

The third page of this report is the misses. They're the most informative numbers in the document.

WHERE WE LANDED WELL

Items inside the margin.

The top five items from this run, ranked by overlap. These reproduce the Pew marginals to within a few percentage points.

TOP-5 SUMMARY

0.934
MEAN OVERLAP

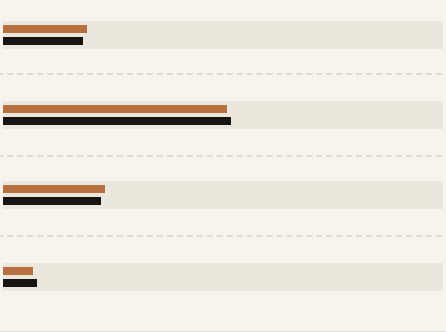
5.1
MEAN MAE (PP)

How to read these. Each card shows the question key, the question as fielded, and the overlap score. Per option, the upper copper bar is **Replism**, the lower black bar is the **Pew ground truth**, and Δ is the signed gap in percentage points.

SM9_A_W163

How well do you think social media helps give a voice to underrepresented groups?

OVERLAP 0.982



	REP	PEW	Δ
Very well	19.0	18.1	+0.9
Somewhat well	51.0	51.7	-0.7
Not too well	23.0	22.4	+0.6
Not at all well	7.0	7.8	-0.8

REPLISM PEW GROUND TRUTH 0 to 100%

SNSUSE_W163

Do you ever use social media sites like Facebook, X (formerly Twitter), or Instagram?

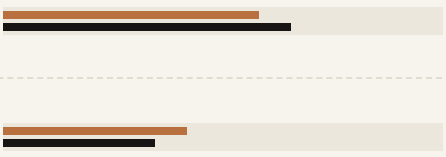
OVERLAP 0.937

	REP	PEW	Δ
Yes, I use social media sites	87.0	81.0	+6.0
No, I do not use social media sites	13.0	19.0	-6.0

TRANSGEND1_W163

Which statement comes closer to your views, even if neither is exactly right?

OVERLAP 0.925



	REP	PEW	Δ
Whether someone is a man or a woman is determined by the sex they were assigned at birth	58.0	65.3	-7.3
Someone can be a man or a woman even if that is different from the sex they were assigned at birth	42.0	34.7	+7.3

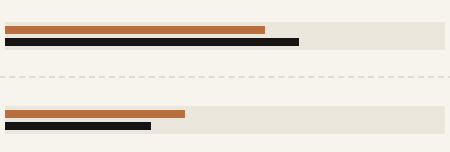
WHERE WE LANDED WELL · CONTINUED

The pattern in the top of the distribution is consistent: two-way questions with clear language, well-balanced base rates, and minimal social-desirability load are where Replism is most reliable today.

TRANGEND1MOD_W163

Which statement comes closer to your views, even if neither is exactly right? (modified wording)

OVERLAP 0.919

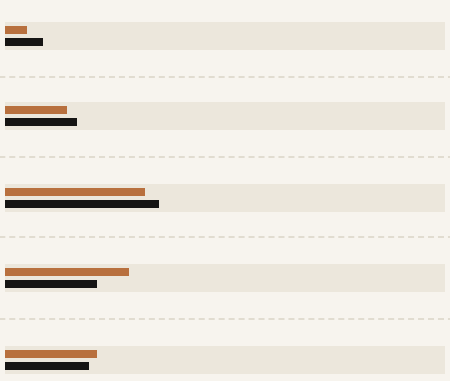


Whether someone is a man or a woman is determined by their sex at birth	REP	59.0	Δ -7.7
	PEW	66.7	
Someone can be a man or a woman even if that is different from their sex at birth	REP	41.0	Δ +7.7
	PEW	33.3	

STATERACEFEEL_G_W163

When you think about the state of race and racial issues in the U.S. these days, how often do you feel uninterested?

OVERLAP 0.908



Extremely often	REP	5.0	Δ -3.6
	PEW	8.6	
Very often	REP	14.0	Δ -2.3
	PEW	16.3	
Sometimes	REP	32.0	Δ -3.1
	PEW	35.1	
Rarely	REP	28.0	Δ +7.1
	PEW	20.9	
Never	REP	21.0	Δ +1.9
	PEW	19.1	

TAKEAWAY · WHAT THE TOP OF THE DISTRIBUTION TELLS US

On clearly worded items, *do you use it*, *which view comes closer*, *how well does this statement describe X*, Replism reproduces the country's actual marginal distribution to within a few points. These are the items where the instrument is already production-grade.

WHERE WE FELL SHORT

Items outside the margin.

The bottom five items from this run, ranked by overlap. None of these are catastrophic. The lowest overlap is *0.737*, but the pattern across them is more useful than any single number.

BOTTOM-5 SUMMARY

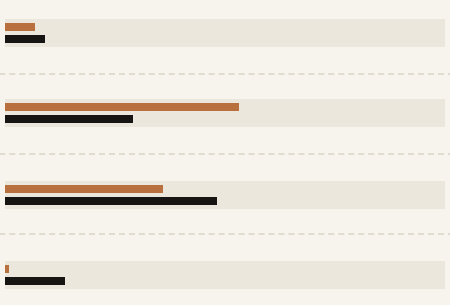
0.764	12.5
MEAN OVERLAP	MEAN MAE (PP)

The shape of the misses. Three of the five worst items are interpretive questions about the legacy of George Floyd's killing, and the other two are scales about social media's importance. The dominant pattern across all five is the same: Replism under-uses the *"Don't know enough to say"* option, and on value-laden race-progress items it commits hard to a position where ground truth shows real ambivalence.

IMPACTPOLICE_W163

Compared with before George Floyd's killing, do you think the relationship between police and Black people in the U.S. today is...

OVERLAP 0.737



	REP	PEW	Δ
Better now than before	7.0	9.3	Δ -2.3
Worse now than before	53.0	28.9	Δ +24.1
About the same as before	36.0	48.0	Δ -12.0
Don't know enough to say	1.0	13.8	Δ -12.8

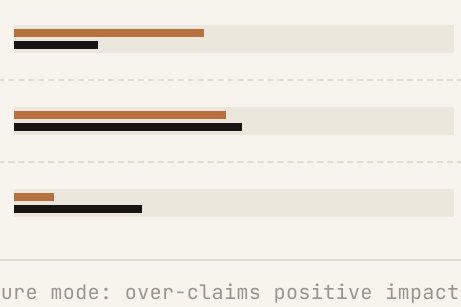
REPLISM · PEW GROUND TRUTH

Failure mode: pessimism overshoot · "Don't know" collapses to zero

IMPLIVES_MOD5YR_W163

Do you think the increased focus on issues of race and racial inequality following George Floyd's killing...

OVERLAP 0.758



	REP	PEW	Δ
Led to changes that have improved the lives of Black people	43.0	18.9	Δ +24.1
Did not lead to changes that have improved the lives of Black people	48.0	51.8	Δ -3.8
Don't know enough to say	9.0	29.3	Δ -20.3

REPLISM · PEW GROUND TRUTH

Failure mode: over-claims positive impact · under-uses "Don't know"

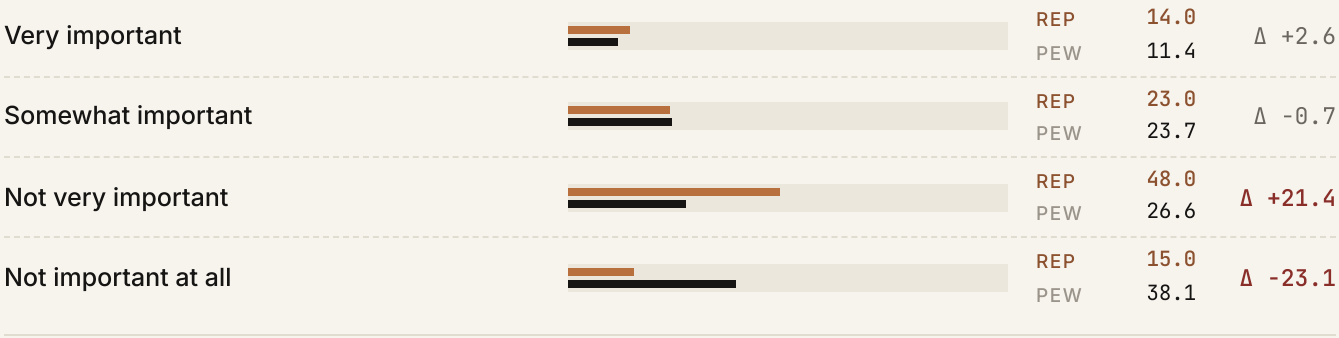
WHERE WE FELL SHORT · CONTINUED

The shared failure mode across these two items is intensity compression on social-media importance scales. The model bunches probability into the middle (“Not very important” / “Somewhat well”) and starves the bottom-end “Not at all” option, even though that’s where a large share of real respondents sit.

SM11_C_W163

How important is social media to you personally for giving you a venue to express your political opinions?

OVERLAP 0.770

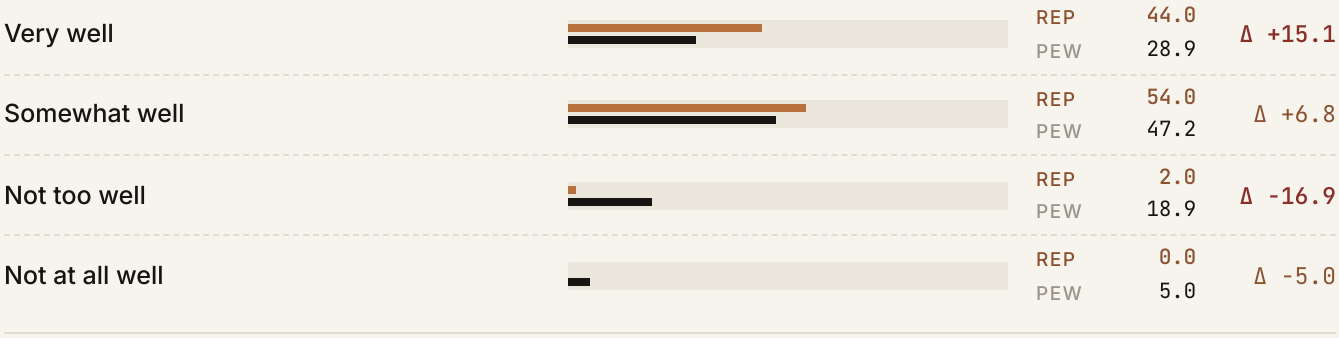


Failure mode: probability mass migrates from “Not at all” into “Not very”

SM9_E_W163

How well does social media make people think they are making a difference when they really aren’t?

OVERLAP 0.774



Failure mode: over-agreement on a value-laden statement · bottom tail collapses

WHERE WE FELL SHORT · PATTERN

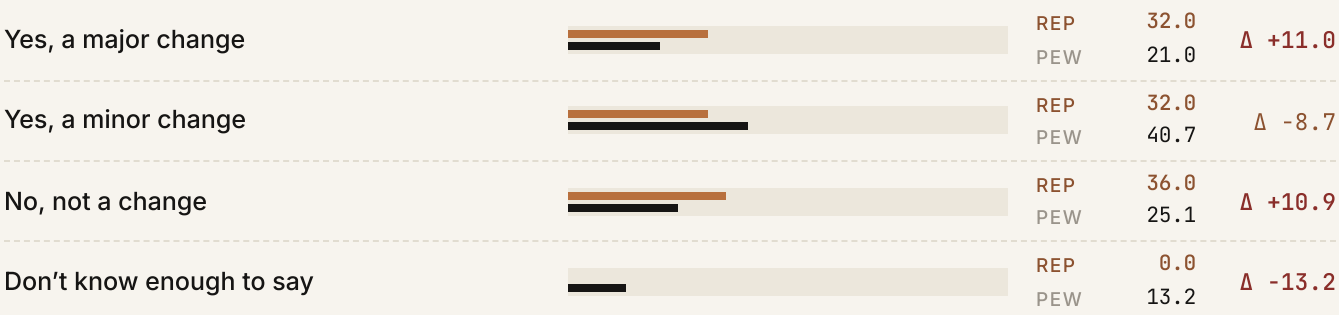
The cleanest illustration of the miss.

The last of the bottom-five items, and the cleanest illustration of the run's dominant failure mode. Ground truth shows roughly one in eight Americans answer *"Don't know enough to say"* when asked to look back on the Floyd-era shift in racial discourse. Replism puts that figure at **zero**, and re-distributes the missing mass across the three confident answers, especially "Yes, a major change" and "No, not a change."

RACECHNG_MOD5YR_W163

Looking back, do you think the increased focus on issues of race and racial inequality following George Floyd's killing represented a change in the way most Americans think about these issues?

OVERLAP
0.779



WHAT THE BOTTOM FIVE TELL US

One shape, two consequences. *"Don't know" suppression* means Replism over-commits where the country is genuinely uncertain. *Intensity compression* on social-media scales means the strongest position, "Not important at all," is systematically under-weighted.

WHAT WE'RE IMPROVING

Reading the misses.

Aggregate overlap is the headline. The misses are the diagnosis.

The bottom of this run shares a signature: **"Don't know" collapses to zero** on retrospective race-progress items, and on social-media importance scales the bottom-end "Not at all" option is starved while the moderate option absorbs the leftover mass.

01

"Don't know" suppression.

Three of the five worst items have a "Don't know enough to say" option that ground truth puts at 13–29%, and that Replism nearly zeroes out.

02

Bottom-tail starvation on importance scales.

On social-media importance items, the "Not important at all" option is the single most under-weighted answer.

03

Over-claim on race-progress retrospectives.

On post-Floyd items, Replism commits hard to a confident view, positive on IMPLIVES, negative on IMPACTPOLICE, where ground truth shows real ambivalence.

Built to be defended, not believed.

REPLISM
EVALUATION REPORT · PEW ATP 163
V0.1 · MAY 2026