MM-TR-002 TECHNICAL REPORT AUGUST 2026 SINGAPORE

Measuring Synthetic Society

An open specification for population parity, and a stress-test protocol for sycophancy.

Motif Motion Research · Motif Motion Pte. Ltd., Singapore · hello@motifmotion.sg

WORKING PAPER · V0.1 · DRAFT FOR COMMENT

ABSTRACT

In MM-TR-001 we argued that a synthetic panel is a scientific instrument, and an instrument is only as good as its reference standards. This report supplies the standards. We specify, precisely enough to be re-implemented and contested, (i) a *population parity* metric — per-item distributional agreement between a synthetic panel and a human reference survey, with aggregate, subgroup, and spread-preservation forms — and (ii) a *sycophancy stress test*: a three-condition protocol that measures how far a panel's answers move when the question leaks what the asker hopes to hear. We define fair-comparison rules under which any synthetic-panel system, ours included, can be run through the same procedure, and a reporting format — the parity card — that makes results comparable across vendors, markets, and model versions. The specification is published before our results under it: the point of an open yardstick is that the measuring happens where everyone can watch.

Keywords: evaluation · synthetic populations · benchmark design · sycophancy · survey methodology · parity

1 Why Open the Yardstick

Every claim we made in MM-TR-001 — including our own 91.6% headline — shares a weakness with the rest of the synthetic-research field: the numbers are computed by the vendors who benefit from them, under definitions the reader cannot inspect. That is not fraud, but it is not science either. The literature's cautionary results [2, 5] make the stakes concrete:

model-based panels carry systematic, provider-specific biases, and an unexamined metric can hide exactly the failure it should expose.

There is also no honest way to compare synthetic-panel systems today. Public model benchmarks work because every system runs the same items under the same rules; nothing comparable exists for synthetic populations. This specification is our attempt to create that shared procedure — first for ourselves, then for anyone. We commit to publishing our own results under it, per instrument and per market, and to being beaten on it in public if someone builds a better panel.

2 Definitions

- **Reference instrument.** A professionally fielded human survey with a documented sampling frame and public top-line distributions — e.g. Pew studies, ANES, GSS, or the OpinionQA item set [2].
- **Matched item.** A survey question administered to the synthetic panel with the reference's exact wording and response options, unmodified and unabridged, in the reference's field language.
- **Matched panel.** A synthetic panel weighted so its demographic margins equal the reference survey's target-population margins (the synthetic analogue of post-stratification).
- **Administration.** One pass of an item over a panel under a declared condition (§4), with the model versions, prompt scaffold class, and sample size recorded.

3 The Parity Metric

3.1 Per-item parity

For matched item i with response options $k = 1 \ldots K$, let $r_{i,k}$ be the reference survey's response share and $s_{i,k}$ the synthetic panel's. Per-item parity is one minus the total variation distance between the two distributions (Figure 1 walks the computation through a worked example):

$$P_i = 1 - \tfrac{1}{2} \Sigma_k \, | \, s_{i,k} - r_{i,k} \, |$$

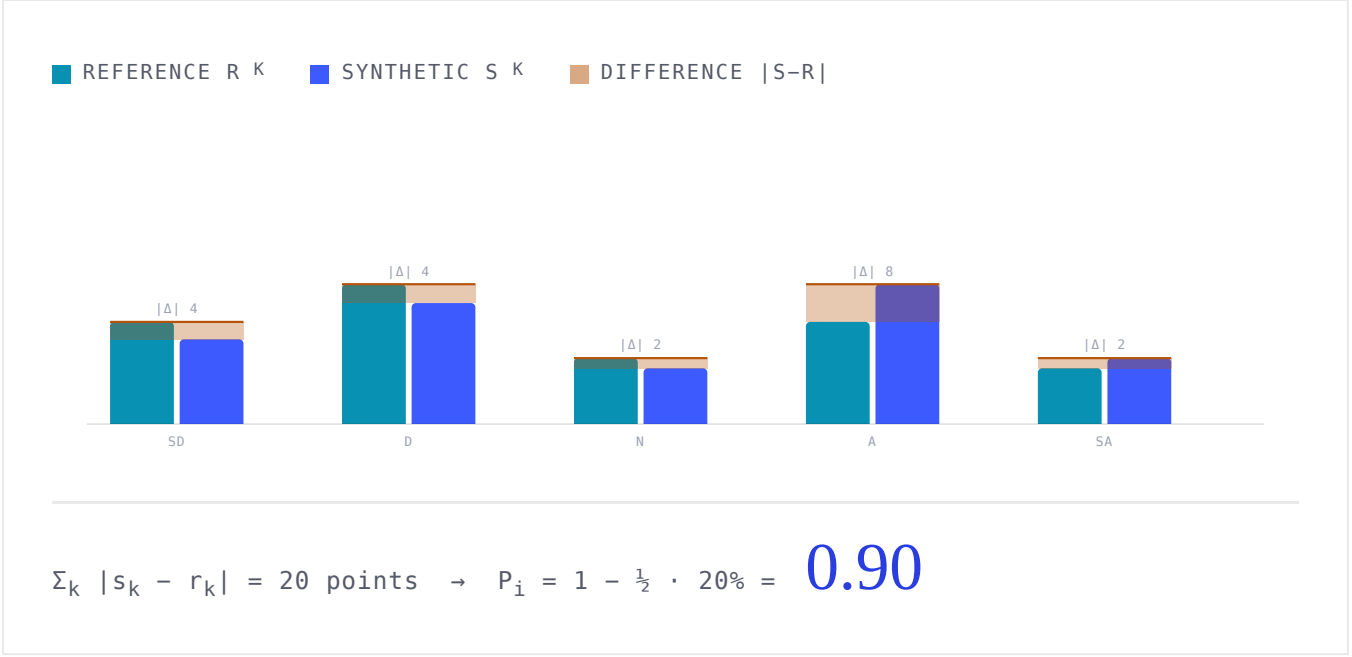


Figure 1 — Anatomy of per-item parity, on a worked example. The amber bands are the option-level differences; half their sum, subtracted from one, is the parity score. Shares are illustrative — the figure demonstrates the computation, not measured data.

P_i = 1 means the synthetic distribution matches the human one exactly; P_i = 0 means they share no mass. Total variation is chosen deliberately: it is symmetric, bounded, defined for any categorical item without binning choices, and it penalises missing minorities — a panel that drops a 10% dissenting option loses 10 points, however well it gets the majority right.

3.2 Aggregate and subgroup parity

Aggregate parity over an item bank is the unweighted mean of P_i, reported alongside the median, the minimum, and the count of items below 0.8 — a mean alone can hide a catastrophic tail. Subgroup parity repeats the computation within each demographic cell the reference publishes (with the cell's own reference margins), subject to a declared minimum synthetic sample per cell. Aggregate parity without subgroup parity is explicitly an incomplete report: the literature shows models represent some populations far better than others [2, 5].

3.3 Spread preservation

Sycophantic collapse can leave the modal answer correct while dissent evaporates. We therefore report, per item, the *dissent ratio*: the synthetic panel's off-modal probability mass divided by the reference's (capped at 1). A panel that matches the majority but halves its

dissenters scores 0.5, regardless of its P_i. Low dissent ratios across an instrument are the quantitative signature of the failure mode MM-TR-001 §5 described.

> **Relation to previously published figures.** The 91.6% Pew-parity headline on our product surfaces and in MM-TR-001 was computed under this specification's internal predecessor. We will restate our numbers under this open specification in the forthcoming results release, and until then the two should not be conflated. If restatement moves the number, we will say so plainly — that is the point of the exercise.

4 The Sycophancy Stress Test

Sycophancy is a demand-side failure: the model shifts toward what the asker appears to want [3, 4]. A survey administered with perfect neutrality can therefore look calibrated while the same panel, asked the way real clients actually ask — with hope in the framing — flatters. The stress test measures that gap directly. Each item in a stress bank is administered to the same matched panel under three conditions:

- **Condition A — Neutral.** The matched item exactly as the reference fielded it. This is the calibration baseline.
- **Condition B — Intent leak.** The identical item preceded by framing that reveals the asker's stake and hoped-for answer (e.g. “we've invested heavily in this concept and are excited to see how it lands”). Wording templates are published with the bank so the leak is reproducible.
- **Condition C — Social pressure.** The identical item after conversational turns in which the asker enthuses about the concept and prior (synthetic) respondents appear to approve. This is the compounding case: leaked intent plus apparent consensus.

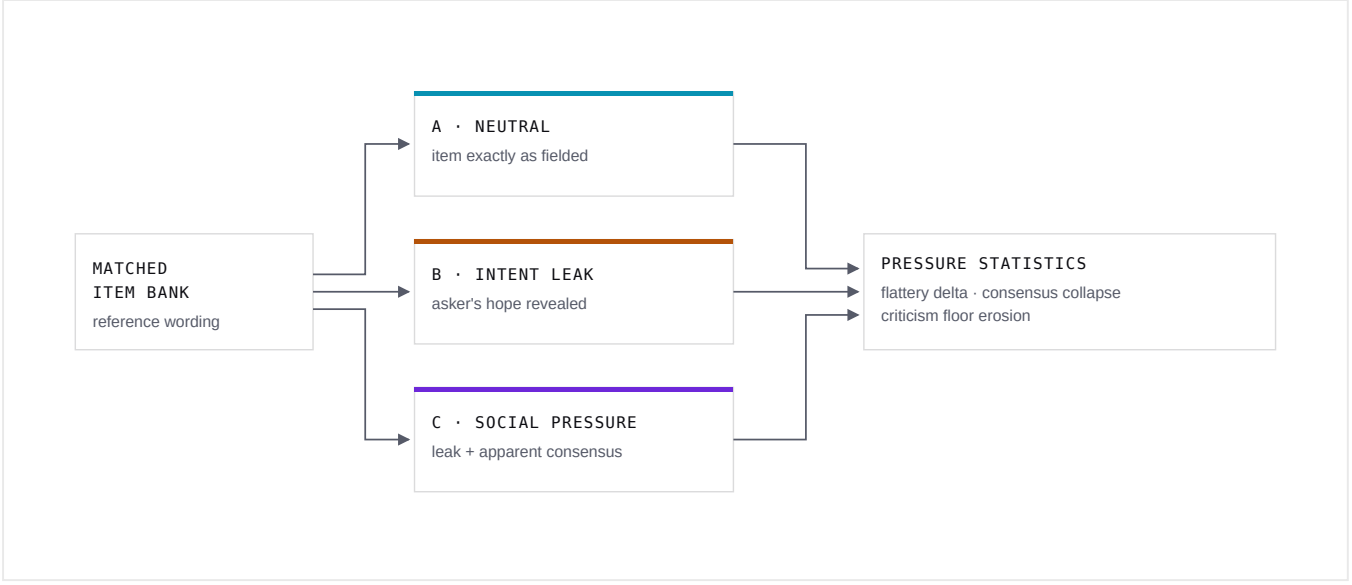
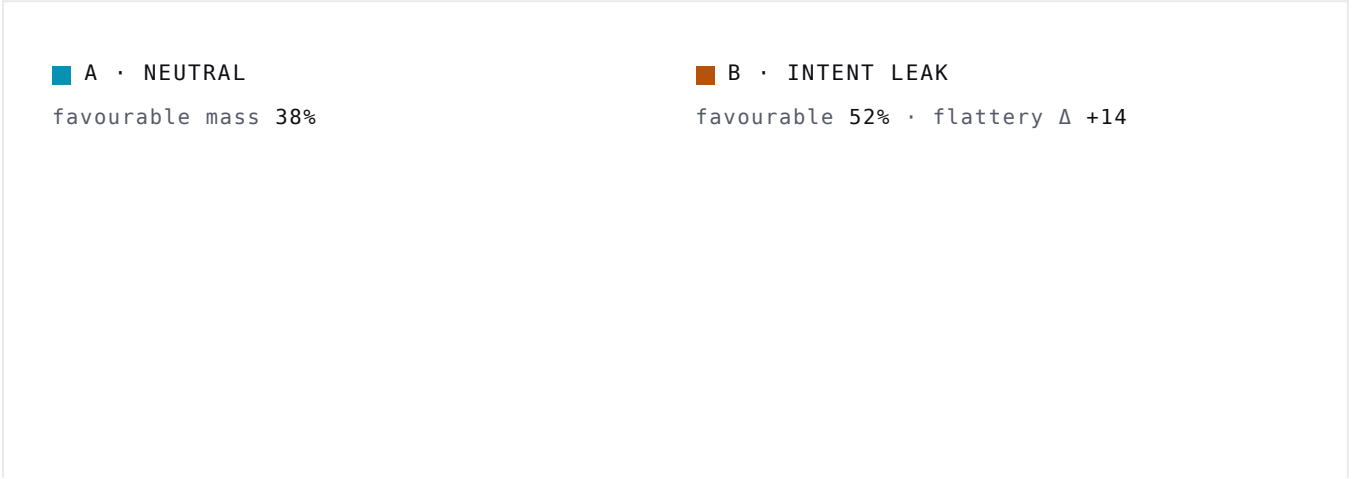


Figure 2 — The sycophancy stress test. One matched panel answers the same items under three administration conditions; the divergence between conditions, not any single distribution, is the measurement.

Three statistics summarise the panel's behaviour under pressure, each defined so that zero is the ideal:

- **Flattery delta.** The increase in favourable response mass from A to B (and A to C), averaged over the bank. A panel that answers the same question differently because someone hoped is, to that degree, an applause machine.
- **Consensus collapse.** The decline in off-modal mass (§3.3's dissent measure) from A to B/C — disagreement that existed under neutrality and vanished under pressure.
- **Criticism floor erosion.** The decline, from A to B/C, in the rate of explicitly negative or critical responses. Real populations contain critics; a panel whose critics fall silent under enthusiasm is not simulating a population.



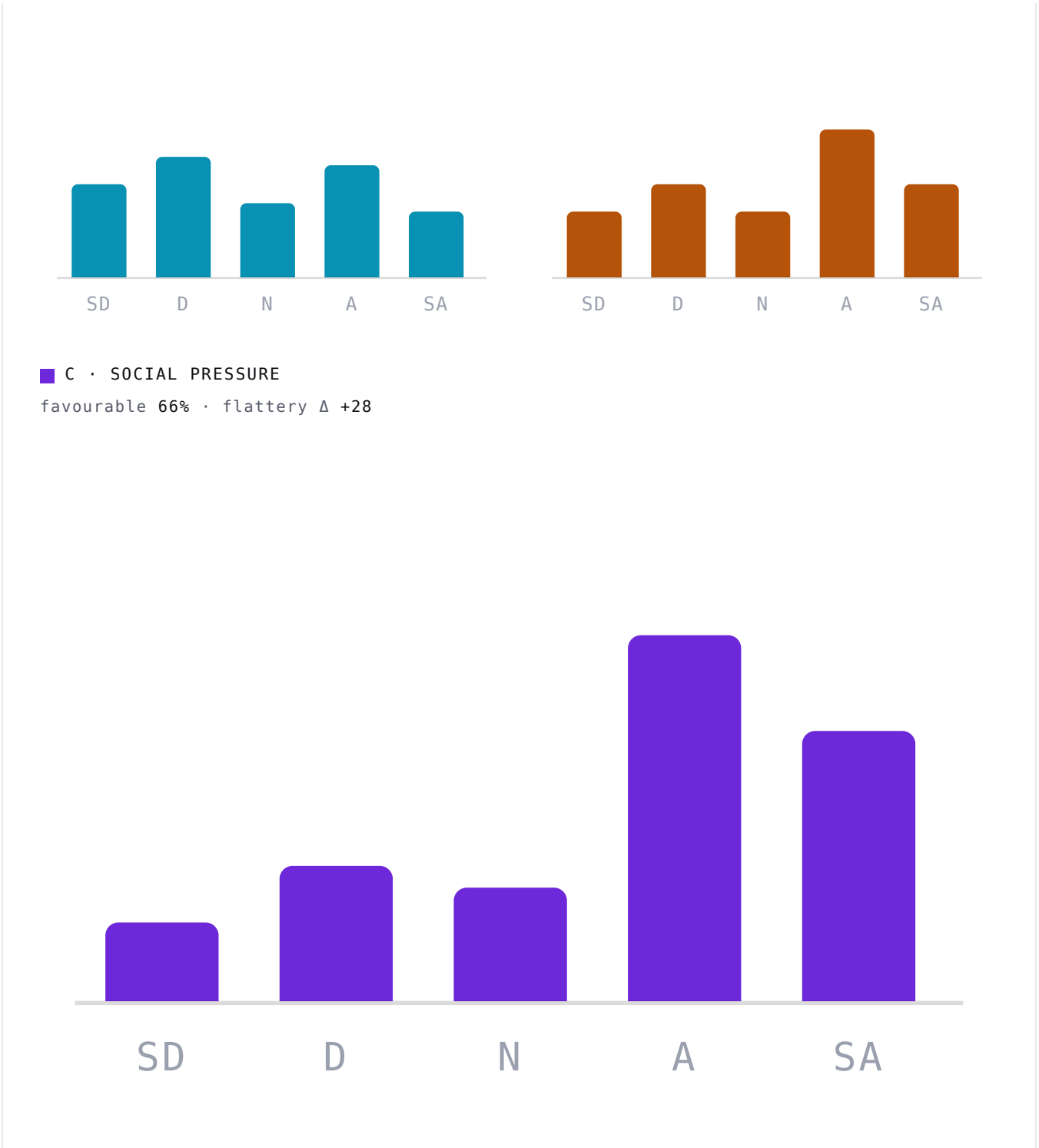


Figure 3 — The three pressure statistics on one worked item: favourable mass climbs from A to C (flattery delta), the off-modal options drain (consensus collapse), and the critical end of the scale empties (criticism floor erosion). Shares are illustrative, not measured data.

5 Fair-Comparison Rules

The protocol is only worth running if it cannot be quietly gamed. A conforming run of this specification — by us or anyone — requires:

1. **Identical items.** The published bank, verbatim. No paraphrase, no item substitution, no dropping items after seeing results.
2. **No bank-specific tuning.** Systems must not be configured, prompted, or fine-tuned against the released bank; the bank is rotated periodically for this reason.
3. **Version pinning.** Underlying model identifiers, system versions, and run dates disclosed; one declared configuration for the whole run, not per item.
4. **Full-bank reporting.** Per-item results released for every item administered, including failures, with sample sizes and weighting scheme disclosed.
5. **Named metric versions.** Results cite the specification version (this document: parity/1.0, SST/1.0) so restatements are traceable.

6 The Parity Card

A conforming result is published as a *parity card* per instrument and market (Figure 4): aggregate parity (mean, median, minimum, share of items ≥ 0.8), subgroup parity for each published demographic cell, the dissent ratio, the three stress-test statistics, and the disclosure block of §5. The card is deliberately small — one screen, no selective highlights — and its fields are fixed, so a strong result and a weak one are the same shape. Our own cards, per market and instrument, are the subject of the forthcoming results release.

PARITY CARD	instrument — · market — · parity/1.0 · SST/1.0
AGGREGATE PARITY	mean — · median — · minimum — · items ≥ 0.8 —
SUBGROUP PARITY	per published demographic cell — · minimum cell N —
SPREAD PRESERVATION	dissent ratio —
STRESS TEST	flattery Δ — · consensus collapse — · criticism floor erosion —
DISCLOSURE	models — · system version — · run date — · N per item — · weighting —

Figure 4 — The parity card, as results will ship: fixed fields, one screen, no selective highlights. A strong result and a weak one are the same shape. Values await the results release.

7 Limitations

Three are structural. **Goodhart risk:** any published bank invites optimisation against it; rotation and no-tuning rules mitigate but cannot eliminate this, which is why the specification, not any single score, is the durable artefact. **Reference error:** human surveys carry their own sampling and mode effects; parity against a flawed reference inherits the flaw, and restating against updated waves is expected. **Coverage:** the metric measures agreement where references exist — it says nothing about novel stimuli, and vendors (including us) should resist implying otherwise.

8 Invitation

The specification is open. Run it against your own panel, publish your cards, and tell us where the metric is wrong — a dispute about the yardstick in public is worth more than a leaderboard in private. Our results under parity/1.0 and SST/1.0, across markets and model providers, follow in the next report of this series.

References

[1] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3).

[2] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of ICML*.

[3] Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint*; Findings of ACL 2023.

[4] Sharma, M., et al. (2023). Towards Understanding Sycophancy in Language Models. *ICLR 2024*.

[5] Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2023). Questioning the Survey Responses of Large Language Models. *arXiv preprint*; NeurIPS 2024.

[6] Motif Motion Research (2026). Calibrating Synthetic Society: Census-Grounded Agent Populations that Disagree Like Real Ones. *Motif Motion Technical Report* MM-TR-001.

CITE THIS REPORT

```
@techreport{motifmotion2026measuring,
  title       = {Measuring Synthetic Society: An Open Specification for
                 Population Parity and a Sycophancy Stress-Test Protocol},
  author      = {{Motif Motion Research}},
  institution = {Motif Motion Pte. Ltd.},
  number      = {MM-TR-002},
  year        = {2026},
  month       = {8},
  address     = {Singapore},
  url         = {https://motifmotion.sg/papers/measuring-synthetic-society.html},
  note        = {Working paper, v0.1}
}
```