Measuring What We Mean: Construct-Identified Inference When Measures Are Cheap

Augusto Gonzalez-Bonorino[1] Daniel J. Martin[2]

August 2026

[1]Department of Economics, Arizona State University; EconLLM Lab.
[2]Perwell (affiliation to be confirmed).

Statement of significance

Scientists rarely observe constructs such as patience or trust directly; they work through measures. Artificial intelligence has made candidate measures cheap, so the scarce task is deciding which measures deserve the interpretations researchers give them, and how far a scientific conclusion moves when different defensible measures are used. We answer with a reporting discipline assembled from three existing traditions: construct validation, multiverse analysis, and partial identification. The researcher declares four things: the construct, a finite menu of candidate measures, a set of pre-stated expectations a credible measure should satisfy—each with a numeric threshold—and the scientific quantity of interest. Measures that satisfy every expectation are admissible; the paper then reports the range of the quantity of interest across admissible measures alongside the unrestricted range across the whole menu. The contrast between the two ranges is itself evidence, and rejected measures are not discarded silently: when they disagree in structured ways, that pattern is evidence that the expectations—or the construct definition behind them—may need revision. Every reported range is conditional on the expectations being correct implications of the construct, an assumption the paper states prominently rather than buries. Under a predeclared within-country education implication, the only admissible cheap trust score is negatively associated with GPS trust (plug-in $\beta = -0.317$). GPS trust itself fails the country-level network, so that association is recovery of a contested criterion, not a refutation of a validated one. A country-cluster band on that singleton crosses zero.

Abstract

Candidate measures of latent constructs have become cheap: survey modules, proxy items, composites, prompted language models, and trained policies can each score the same construct, and different scores can move conclusions. This paper develops *construct-identified inference*, a reporting discipline for that abundance. The researcher declares a menu of candidate measures and a validity argument—pre-stated convergent and discriminant expectations with researcher-owned thresholds—which screens the menu into an admissible set. The headline object is the image of the downstream estimand over admissible measures, always reported alongside the unrestricted image over the full menu; the contrast between the two separates homogeneity created by the screen from homogeneity that survives it. Screening is auditable under a stated freeze: networks and thresholds are hash-pinned before slacks are inspected. This study was not preregistered. Sampling fragility induced by data-dependent screening is reported as an additive uncertainty band, not folded into a confidence set. Structured disagreement among rejected measures is read as evidence about the validity argument rather than discarded, and every conclusion is conditional on the restrictions being correct testable implications of the construct. The framework ships in an open, score-agnostic package and is applied to patience and trust, combining human survey measures with open-weight language-model scores at two resolutions. Neither application yields a preferred measure; the reported disagreements are what the composition is for.

Keywords: construct validity; criterion-related validity; partial identification; multiverse; AI measurement

Plain-language abstract

Scientists rarely observe things like patience or trust directly; they observe scores—survey answers, composite indexes, or, increasingly, outputs of language models—and different measures can lead to different conclusions. Artificial intelligence has made candidate measures cheap, so the hard problem is deciding which measures are trustworthy and which conclusions survive that decision. This paper proposes a reporting rule: researchers declare a menu of candidate measures and a set of pre-stated expectations a credible measure should satisfy (for example, a patience measure should rise with education and should not collapse into risk-taking); measures that satisfy every expectation are admissible, and the paper reports the range of the scientific result across those admissible measures rather than a single preferred score. If no measure survives, that emptiness is a finding, not a failure. The rule is implemented in a free, open-source software package and demonstrated on international survey data on patience and trust, including scores produced by two open-weight language models. The demonstration shows the rule can fail—the education screen and GPS-trust recovery pull apart: an unrestricted menu of patience measures sign-switches on a development question; the canonical trust item behaves more like a patience measure than a trust measure; and the only cheap trust score that clears the education bar is negatively associated with GPS trust, which itself fails the country-level network, so that association recovers a contested criterion rather than refuting a validated one. The paper's conclusion is that disagreement across measures, models, and resolutions is the measurement result—information to be reported, not noise to be hidden.

1 The inferential problem when measures become cheap

1.1 Science is a pipeline

Empirical science is a pipeline. Hypothesis generation proposes constructs and candidate operationalizations; measurement generation maps inputs into scores; construct validation asks whether those scores support the interpretation they are given; estimation uses the scores in downstream analyses; and a conclusion stage reports what has been learned. Each stage has its own tools and its own failure modes, and the stages compose: a conclusion is only as strong as the weakest link in the chain (Figure 1). For most of the history of empirical social science, the binding constraint sat at the measurement stage: producing a new measure was costly, and validation effort concentrated on a few carefully constructed instruments. Artificial intelligence relaxes that constraint. Prompted language models, log-probability elicitation, embeddings, and trained policies can each produce a candidate score for the same construct at low marginal cost [29, 40], and the supply of candidate scores grows quickly. What does not grow is the evidence that any particular score column measures the construct. In particular, a standard error computed conditional on a chosen score column—however small—cannot establish that the column measures the construct: sampling variability and construct validity are different objects, and no amount of the former substitutes for the latter. When measures become cheap, validation becomes the scarce stage of the pipeline.

1.2 The measured failure mode

The change is not hypothetical; practice has been audited. Baumann et al. replicate annotation tasks from published social-science studies and show that plausible LLM prompt choices produce incorrect downstream conclusions for roughly 31% of hypotheses [3]. Desai, Card, and Jacobs audit LLM-based measurement in flagship social-science journals and find validation practices inconsistent and often limited relative to the centrality of the measurements [14]. Kim et al. document that LLM errors are correlated far beyond chance—same-wrong-answer agreement of 42% against a 13% chance baseline, with better models *more* correlated, not less [34]. Those audits document undeclared menus and correlated generator error; they are not, as reported, failed restrictions against a declared network. Two implications follow. First, agreement across models is weak evidence of validity: models that share training pipelines can share mistakes, and consensus may reflect correlated error rather than construct-relevant signal. Second, engineering quality is not a substitute for validation. Yin, Vu, and Persico show that LLM-based occupational-exposure scores—inputs to a large applied literature—are unstable across model choices [56]. The scale is measurable: in Hall's hand-coded classification of the PolMeth 2026 program, 26% of presentations were core generative-AI, and annotation plus measurement accounted for about 60% of those [25]. Measurement generation has arrived in the pipeline; validation practice has not kept pace.

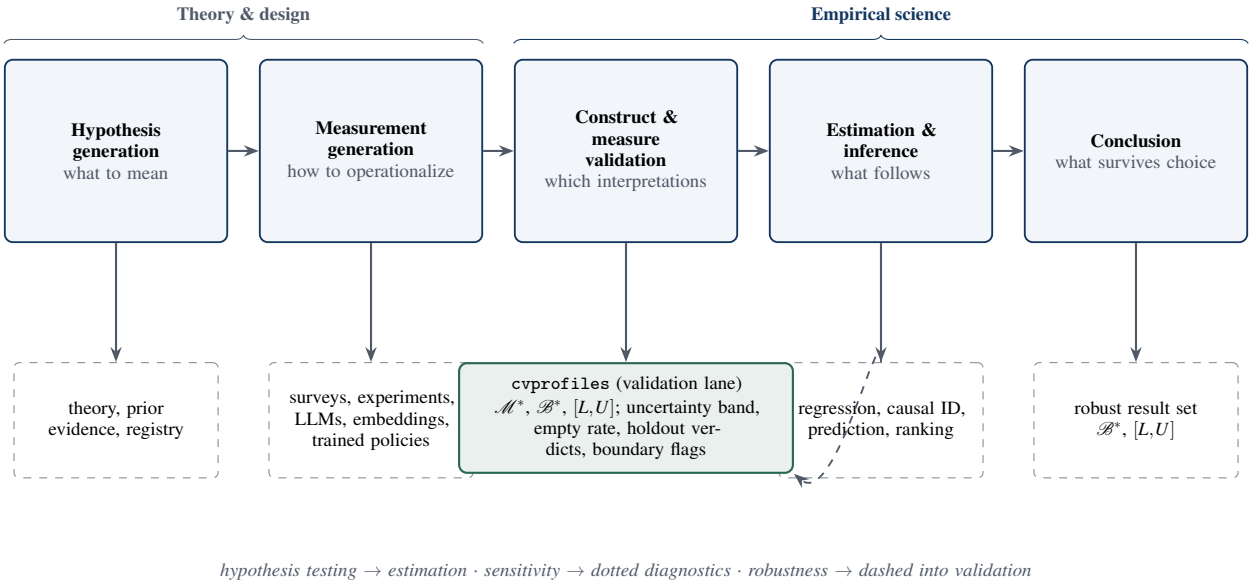


Figure 1: A systems view of empirical knowledge production. AI primarily expands the measurement-generation stage; this paper disciplines the bridge from construct validation to estimation and conclusions. Solid arrows denote artifacts, dashed denote researcher-owned commitments, and the diagnostics that monitor the validation lane are reported separately from the headline outputs.

1.3 The missing composition

The audits above share a feature: each documents a failure inside a single stage of the pipeline. What is missing is the composition—an account of how measurement, validation, and estimation combine into a scientific conclusion, and of what a robustness claim means once measures are cheap. Without such an account, two reflexes dominate. One declares a result fragile because it shifts under any new operationalization, treating every plausible score column as equally entitled to enter the analysis. The other declares a result robust because a single operationalization was retained. Both reflexes confuse the menu of candidate measures with the evidence about the menu. The claim this paper works from is stated plainly. A scientific result should not be declared heterogeneous because it changes under operationalizations that fail a declared validity argument, nor homogeneous because a single operationalization passes one. Two images of the downstream estimand are therefore reported side by side: the unrestricted result set over the full menu, and the validity-argument-conditional set $\mathscr{B}^*(\mathscr{R}) = \{\beta(m) : m \in \mathscr{M}^*(\mathscr{R})\}$ over the survivors; their contrast is evidence about the measures and, when rejection is structured, about the argument itself. The discipline is deliberately one-directional in the confirmatory lane: validity evidence screens measures. Classical construct validation is reciprocal—anomalous patterns among rejected measures can motivate revising the construct or the network—and we state that discovery channel explicitly in Appendix D rather than collapsing it into the confirmatory screen. All conclusions are conditional on the maintained assumption that each restriction is a correct testable implication of the construct (Section 3.2). The rest of this section makes the objects precise and states the contributions.

1.4 Four objects

The framework needs four objects. A *construct* C is the scientific concept of interest, defined by its intended interpretation rather than by any single column. A *menu* $\mathscr{M}$ is the finite set of candidate measures the researcher declares. A *validity argument* $\mathscr{R}(C)$ encodes what a measure of C should do: convergent and discriminant implications derived from the construct's nomological role, written as moment inequalities $\mathbb{E}[g_r(m,V;\tau_r)] \geq 0$ over auxiliary variables V, with researcher-owned thresholds τ_r. A *downstream estimand* $\beta(m)$ is the scientific quantity computed when measure m is used. The validity argument screens the menu: a measure survives when its sample slack $s_r(m)$ clears every restriction,

$$\mathscr{M}^* \;=\; \big\{m \in \mathscr{M} : s_r(m) \geq 0 \text{ for every } r \in \mathscr{R}(C)\big\}, \tag{1}$$

and the surviving set—not a single retained operationalization—is the object of inference; where the dependence matters we write $\mathscr{M}^*(\mathscr{R})$ to stress that the survivors are a function of the declared argument, never a theory-independent object. Remaining measurement choice is then propagated into the result: the robust result set is the image of the estimand over survivors,

$$\mathscr{B}^* \;=\; \big\{\beta(m) : m \in \mathscr{M}^*\big\}, \tag{2}$$

summarized for a scalar estimand by its range,

$$[L,U] \;=\; \Big[\min_{m\in\mathscr{M}^*}\beta(m),\; \max_{m\in\mathscr{M}^*}\beta(m)\Big]. \tag{3}$$

Equations (1)–(3) define the admissible set, the robust result set, and the headline range. Validity here is whether a score warrants the intended interpretation of C, operationalized as nonnegative slack on a declared $\mathscr{R}(C)$; robustness is whether β survives remaining measurement choice inside $\mathscr{M}^*$; the uncertainty band of Section 3.5 is selection uncertainty, not a confidence interval. The construct-identified result set is deliberately conditional—on the construct, the menu, the validity argument, the thresholds, and the sampling frame. Change any component and $\mathscr{M}^*$, $\mathscr{B}^*$, and $[L,U]$ are different objects; every component must therefore be declared and auditable rather than inferred from the data. Two distinct sensitivities follow: sampling sensitivity at fixed $\mathscr{R}$, quantified by the diagnostics of Section 3.5, and sensitivity of $\mathscr{M}^*(\mathscr{R})$ and $\mathscr{B}^*(\mathscr{R})$ across substantively defensible validity arguments $\mathscr{R}_1, \mathscr{R}_2, \ldots$, which is not sampling uncertainty and does not shrink under resampling. The second layer is executed for the powered application in Appendix G.

1.5 Four contributions

The contributions follow in deliberately conservative order. *(i) Validity-constrained measurement robustness.* $\mathscr{B}^*$ is the image of the estimand over measures that survive a declared validity argument, not an unrestricted multiverse; fragility and robustness are therefore attributed to the right cause. *(ii) A selection rule that can fail.* Leftover restrictions never vote on admission. On the frozen cell scores, construct-guided selection is compared with best-β, best-anchor, and size-matched random subsets (Table 5). At cell resolution the leftover restriction is a rank transform of the selecting education correlation, so the patience comparison is weak; the trust comparison is not: best-β picks a rejected composite, construct-guided picks Phi. *(iii) Separation of headline from diagnostics.* The headline range is reported separately from additive sampling diagnostics: an uncertainty band (selection uncertainty, not a confidence set), an empty-replicate rate, and boundary attribution. *(iv) An auditable, score-agnostic implementation and a two-resolution application.* The method ships in the released package `cvprofiles` 3.0.2 [23]. The confirmatory application places human measures of patience and of trust under one contract at two resolutions, with leftover restrictions, additive sampling diagnostics, and two nested special cases computed on the same objects.

1.6 Roadmap

The remainder of the paper locks two commitments. Novelty is claimed only in the operational composition of three existing traditions, in a selection rule that can fail, and in a released implementation. The application is evidence about two constructs at two resolutions, not a general standard journals should require and not a measure of culture. Section 2 situates the argument among three traditions; Section 3 formalizes the objects, including the empty-network and Campbell–Fiske special cases; Section 4 documents the released implementation; Section 5 reports the two-resolution application; Section 6 states what the framework does and does not claim; Section 7 concludes.

2 Three traditions, one missing layer

2.1 Construct validity supplies the restrictions

Construct validity is the oldest and most mature of the three traditions. Cronbach and Meehl introduced the nomological network: a construct earns its meaning through the laws and observable implications in which it participates, and validation consists of testing those implications [11]. Campbell and Fiske sharpened the empirical content into convergent and discriminant criteria, operationalized through the multitrait–multimethod matrix [8]. Messick and, later,

Kane reframed validity as the defensibility of score interpretation and use—a property of interpretations and uses, not of instruments [42, 32]. Borsboom and colleagues added a realist caution: a measure is valid for a construct only if the construct exists and causally produces the score [6]. This paper takes the observable-implication strand [11, 8]. Kane's view is the pair $(C, \beta(m))$, not an extra screen on m; Borsboom's causal-production standard is not required for the plug-in admissible set, so an empty $\mathscr{M}^*$ is a failed nomological test, not a failed existence claim. Computational social science has begun importing this vocabulary. Li et al. formalize the proxy presumption—treating embeddings as measures without evidence—and propose a construct-validity protocol with discriminant, incremental, and predictive checks [37]. Halterman and Keith stage validation for LLM-based coding systems [26]. Licht et al. show that the elicitation protocol, such as pairwise comparison over direct scoring, materially changes what LLM-based measures record [38]. These contributions share a unit: each validates a *procedure*—a scoring method, an item, an instrument—against a criterion. Campbell and Fiske already place several measures in one matrix; the matrix is evidence about each cell. What that literature does not do is take the survivors as the domain of a downstream estimand and report $\mathscr{B}^* = \{\beta(m) : m \in \mathscr{M}^*\}$.

2.2 Multiverses expose operationalization uncertainty but do not adjudicate meaning

A second tradition quantifies how conclusions depend on analytic choices. Multiverse analysis enumerates a large space of defensible specifications and reports the distribution of results across it [52]; specification-curve analysis supplies a reporting discipline for the same idea [50]. At the measurement layer the logic is striking: Hanel and Zarzeczna enumerate 262,143 operationalizations of well-being and show that conclusions about the well-being of religious groups depend on which operationalization one examines [27]. The multiverse is an exploratory surface: it maps where conclusions live under alternative choices. It does not, by itself, adjudicate which choices carry the intended interpretation. Del Giudice and Gangestad make the point sharply: candidate measures differ in validity and reliability, and treating them as interchangeable branches of a multiverse can itself mislead [12]. Unrestricted and screened result sets therefore differ in evidential status, and neither replaces the other: the unrestricted image describes heterogeneity across plausible operationalizations, the screened image heterogeneity among operationalizations the argument licenses. The multiverse tradition also assigns the full menu a stronger role than exploration: structured heterogeneity across operationalizations can itself indicate that the construct or its network is misspecified—Cronbach and Meehl explicitly treated anomalous patterns as potentially requiring revision or subdivision of the construct [11]—and Section 5.4 reads a coherent rejected family exactly that way. The relevant robustness set is nonetheless not the image of the estimand over every plausible operationalization; it is the image over the measures that survive a declared validity argument. The economics tradition reached the same fork a generation earlier. Extreme-bounds analysis ranged the estimand over every defensible specification, with no admissibility test separating credible from incredible choices [36]; the resulting verdict—"everything is fragile"—treated that breadth as the collapse of specification robustness [48]. The forking-paths critique generalized the mechanism: with enough analytic freedom, some defensible path produces any desired conclusion [21]. This paper's screen is the admissibility test EBA lacked, moved from specifications to measurement functions: the unrestricted multiverse over $\mathscr{M}$ is the named special case, and robustness is judged over the survivors $\mathscr{M}^*$, with the validity argument itself held-out falsifiable (Section 3.4).

2.3 Partial identification supplies the set logic, not the validity argument

The third tradition supplies the formal set logic. Partial identification asks what can be learned about a parameter from data and maintained assumptions alone, reporting identified sets rather than point estimates [41, 44]; confidence sets for partially identified parameters extend the logic to sampling uncertainty [31, 53]. The connection to measurement is recent and explicit. In the 2026 NBER Summer Institute Methods Lectures, *Estimation and Inference with AI-Generated Data*, Dell and Rambachan frame the nomological network as partial identification under theoretically motivated moment restrictions—Dell on measurement and construct validity, Rambachan on multiple measurements and data fusion—and explicitly call for principled, widely accepted methods for construct validity [13]. The nearest formal neighbor is Chen, Rambachan, and Tamer, whose restrictions are benchmark-informed bounds on error correlation for a downstream parameter [10]. The present framework differs in the object of inference: our restrictions are menu-level nomological moment inequalities, the surviving set $\mathscr{M}^*$ is the object, and the selection rule is held-out falsifiable. Other recent work occupies adjacent territory without a surviving set: Ogut and Yin derive curvature bounds from multiple nonlinear measurements [46]; Messing shows that naive standard errors in LLM evaluation pipelines can be 40–60% too small—a diagnosis about evaluation scores, not construct measures [43]; Canen and Enamorado correct regressor

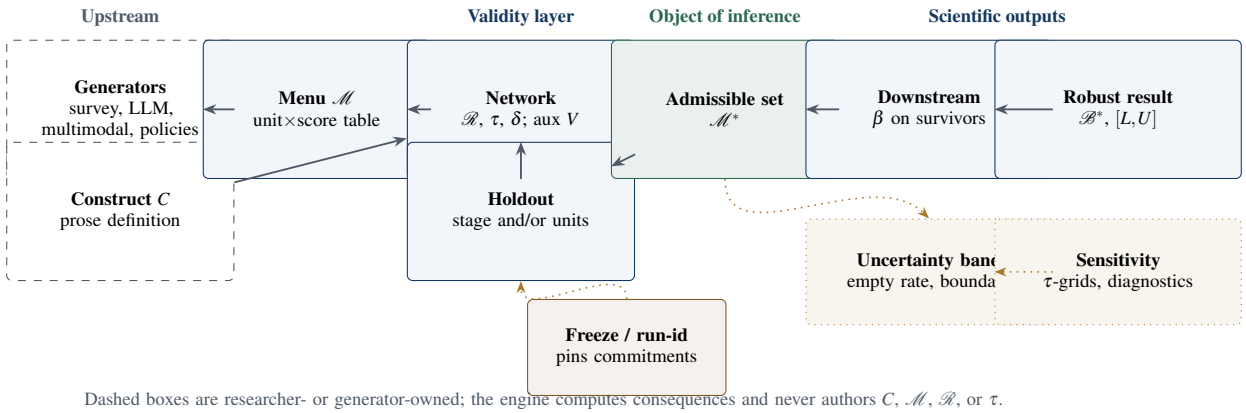


Figure 2: Core objects and epistemic ownership of construct-identified inference. The admissible set $\mathscr{M}^*$ is the object of inference; $\mathscr{B}^*$ and $[L,U]$ are downstream scientific outputs computed only over survivors; sampling and sensitivity diagnostics live in a separate dotted lane; and a freeze/run-id pin sits between researcher commitments and engine compute.

measurement error by split-sample procedures—a correction, not a selection [9]. None performs menu-level validation with a surviving set and a falsifiable selection rule.

2.4 AI makes the old problem large

The fourth development is not a tradition but a shock. AI measurement generators—prompted elicitation, log-probability scoring, persona steering [33], preference fine-tuning [47, 1, 15], and synthetic-agent pipelines such as SCA [24]—expand the supply of candidate scores at low marginal cost. Each technology enters $\mathscr{M}$ only as a map on the same declared units; pipelines that replace the sample, including some SCA-style designs, are not menu expansions under Definitions A.1–A.3. Engineering sophistication is not construct validity: a fluent generator can produce a score column that fails every nomological implication of the construct it purports to measure. The correct discipline is to keep generators and validators separate. Generators expand the menu; they do not adjudicate it. That separation, long standard for conventional measures, becomes the central discipline when measures are cheap. AI does not create the measurement problem; it makes an old problem large.

2.5 Synthesis

Each tradition contributes one piece. Psychometrics supplies a validity-evidence architecture from which restrictions can be derived; the restrictions used here draw on evidence based on relations to other variables, within the broader ecosystem of internal-structure, reliability, invariance, and response-process evidence. The multiverse tradition supplies the demand that conclusions be reported across choices rather than under a single one. Partial identification supplies the set logic in which a surviving set and its image are well-defined. The composition those traditions do not jointly supply is the map $\mathscr{R}(C) \to \mathscr{M}^*(\mathscr{R}) \to \mathscr{B}^*(\mathscr{R})$, drawn with a return arrow: observed disagreement among measures can motivate revision of C or $\mathscr{R}$ for future studies. A holdout that makes screening prospectively falsifiable, and a freeze that makes it auditable, are protocol commitments; they are not implied by the map. That composition—the ordering of established ideas rather than a new taxonomy—is the contribution of this paper.

3 Construct-identified inference

This section fixes the objects, the screen, the falsification designs, and the sampling diagnostics.

3.1 Primitives and notation

Every object of the framework has a formal counterpart; Table 1 collects them and Figure 2 assembles them into the workflow. A *construct* C is the scientific concept of interest, defined in prose by its intended interpretation and

Table 1: Objects of construct-identified inference. The owner column records who authors each object: the researcher or theory, the measurement generator jointly with the researcher, or the engine itself. The engine computes consequences of commitments; it does not author the construct, the menu, the restrictions, or the thresholds.

Object	Definition	Scientific owner
C	Construct defined in prose by its intended interpretation, independent of any score column	researcher / theory
m	Measurement function $m : \mathscr{X} \to \mathbb{R}$, or a declared scalarization; maps inputs to a unit-level score	generator + researcher
$\mathscr{M} = \{m_1, \ldots, m_J\}$	Finite declared menu of candidate measures	researcher / theory
V	Auxiliaries: observed variables used to evaluate validity implications	researcher / theory
$\mathscr{R}(C)$; τ_r	Nomological restrictions with substantive thresholds	researcher / theory
$\mathscr{M}^*$	Admissible set: measures passing every restriction up to tolerance δ	computed
$\beta(m)$	Downstream target functional evaluated at measure m	researcher / theory
$\mathscr{B}^*$; $[L,U]$	Image of β over $\mathscr{M}^*$; min–max scalar summary	computed
$G(D,C)$	Measurement-generation technology: evidence D and construct C produce candidate measures	generator + researcher

independent of any score column. A *measurement function* $m : \mathscr{X} \to \mathbb{R}$ maps available inputs to a unit-level score, or to a vector followed by a *declared* scalarization. The *menu* $\mathscr{M} = \{m_1, \ldots, m_J\}$ is the finite set of candidates declared by the researcher. *Auxiliaries* V are observed variables used to evaluate validity implications; they are not the construct. Observed does not mean error-free: when V is itself a score for another construct, its measurement properties condition the reading of any restriction built on it. The *nomological network* $\mathscr{R}(C)$ encodes those implications as moment restrictions with substantive thresholds τ_r, following the classical validation tradition [11, 8] and its recent partial-identification reading [13]. We reserve *nomological network* for the broader theory from which $\mathscr{R}(C)$ is derived; $\mathscr{R}(C)$ itself is the finite, operationalized set of thresholded implications actually screened. A failed restriction can therefore mean either that m is a poor measure or that the network supplied the wrong implication—the screen cannot by itself distinguish the two. The admissible set $\mathscr{M}^*$ retains the measures that pass every restriction; $\beta(m)$ is the downstream target; $\mathscr{B}^*$ is its image over $\mathscr{M}^*$, with scalar summary $[L,U]$; and $G(D,C)$ denotes a measurement-generation technology producing candidate measures from evidence D and construct C. The owner column of Table 1 states the division of labor: commitments are authored upstream; the engine computes their consequences.

3.2 Restrictions as observable implications

A validity restriction must be a decision rule with a prespecified direction. We write each restriction $r \in \mathscr{R}(C)$ as a population moment inequality,

$$r: \qquad \mathbb{E}\big[g_r\big(m(X), V; \tau_r\big)\big] \geq 0, \tag{4}$$

where τ_r is the *substantive threshold*: a researcher-owned commitment justified by theory, prior literature, or a preregistered tolerance, and *not* a parameter estimated by the admissibility procedure. Weak inequalities are the normal form of one-sided validity content ("at least this association," "no more than this contamination"). A numeric gloss fixes the reading: $\tau_{\text{conv}} = 0.39$ means that any measure whose population correlation with the convergent criterion falls below 0.39 fails the declared admissibility gate under $\mathscr{R}(C)$, regardless of how well it predicts Y. Let $s_r(m)$ denote the sample slack of restriction r on measure m (positive when satisfied). For a declared tolerance $\delta \geq 0$, the admissible set is

$$\mathscr{M}^*(C; \mathscr{M}, \mathscr{R}, \delta) = \big\{m \in \mathscr{M} : \; s_r(m) \geq -\delta \;\; \forall r \in \mathscr{R}(C)\big\}. \tag{5}$$

Passing is conditional evidence: $m \in \mathscr{M}^*$ means m is consistent with the declared interpretation—not that m is the construct, that the menu is exhaustive, or that m is valid for uses the tests do not represent. An empty admissible set is a valid finding: it rejects the pair $(\mathscr{M}, \mathscr{R})$ as an operationalization of C, not C itself. A poor restriction can reject a useful measure; a weak network can admit a poor one; and a misspecified network can manufacture apparent homogeneity by excluding a coherent family of counterexamples. The force of the screen is therefore conditional on the restrictions being correct implications of the construct (Assumption A.2): pre-specification fixes researcher degrees of freedom, but it does not by itself establish that a restriction is substantively necessary for the intended interpretation. The framework disciplines these commitments by making them explicit and auditable—it does not make them disappear.

Because every downstream claim inherits it, the maintained assumption is stated formally in the main text rather than left to the appendix:

Assumption 3.1 (Restriction validity)**.** *Each g_r is a correct testable implication of the construct definition and its auxiliaries V; restrictions are researcher-owned and falsifiable by data. If a restriction fails in data, that failure is prima facie evidence against the measure—but the strength of prior evidence behind the restriction determines how much weight any conclusion conditional on it can bear.*

The screen is deliberately not a classical joint hypothesis test: there is no likelihood, no null distribution, and no prior; the engine computes transparent moments against declared thresholds and reports which measures clear them. The scientific content is the intersection of the restrictions, not an aggregate statistic. Concrete tests based on relations to other variables—convergent, discriminant, criterion-related, and broader nomological expectations—fit this form [37, 26], in contrast to restrictions that bound measurement-error correlation structure [10].

3.3 From admissible measures to scientific conclusions

Researchers usually care about an effect, association, ranking, or forecast rather than the score itself; let $\beta(m)$ denote the result when measure m is used. Once the admissible set is defined, the conclusions licensed under the validity argument are the image

$$\mathscr{B}^*(C;\mathscr{M},\mathscr{R}) = \{\beta(m) : m \in \mathscr{M}^*\}, \tag{6}$$

with the transparent scalar summary

$$[L,U] = \Big[\min_{m\in\mathscr{M}^*}\beta(m),\ \max_{m\in\mathscr{M}^*}\beta(m)\Big], \tag{7}$$

defined whenever $\mathscr{M}^* \neq \emptyset$.

The contrast with an unrestricted multiverse makes the ordering concrete [50]. If $\mathscr{R} = \emptyset$, then $\mathscr{M}^* = \mathscr{M}$ and $[L,U]$ is the specification-curve range over the full menu: every candidate is entitled to speak, and remaining disagreement is reported rather than screened. That object is a named special case of (7), not a missing network. Suppose five candidate measures deliver β values 0.10, 0.12, 0.35, -0.04, and 0.38. Reported over the full menu, the conclusion ranges from -0.04 to 0.38: sign-switching, and the immediate question is whether the heterogeneity in β is unstructured noise or systematically associated with measure characteristics. If the validity argument retains only the two measures with $\beta(m) = 0.35$ and $\beta(m) = 0.38$, then conditional on $\mathscr{R}$ the construct-identified range $[0.35, 0.38]$ is narrow; $\mathscr{R}$ may be exactly what selected the homogeneous subset, which is why the unrestricted range is reported alongside. If all five survive, every candidate meets $\mathscr{R}(C)$ but admissible measures disagree on β, so no single estimand conclusion survives measurement choice within the declared menu. The multiverse domain is $\mathscr{M}$; the construct-identified domain is $\mathscr{M}^*$. $\mathscr{B}^*$ is not an unconditional identified set for the latent construct; it is the set of results produced by measures that survive a declared validity argument [41, 44]. The unrestricted image over $\mathscr{M}$ and the conditional image over $\mathscr{M}^*$ answer different questions; where both are computable this paper reports both, and the contrast between them is informative rather than a defect.

The same equations recover two classical objects. If $\mathscr{M}^*$ is treated as known, endpoint inference for $[L,U]$ is ordinary inference for extrema over a fixed finite set, and a Bonferroni union of intervals is a conservative cross-check in the spirit of Imbens and Manski [31]. If each g_r is a Chen–Rambachan–Tamer restriction on error correlation [10], (4)–(7) are analogous to their identified set; a finite-menu image is not their identified set without an explicit mapping. Campbell and Fiske's monotrait–heteromethod and heterotrait–monomethod cells [8] are likewise instances of (4): a convergent floor τ_{conv} is a corr_min restriction against a partner column, and a discriminant ceiling τ_{disc} is a corr_zero restriction. The paper computes both special cases on frozen scores in Section 5.3.

3.4 Falsifying the validity rule

A validity framework can become circular: inspect available evidence, write restrictions after seeing which columns would pass, and report that they pass. The two designs below test different claims about a *frozen* contract; they do not constrain how $\mathscr{R}$ was written, and they test the predictive content of that contract, not its unique truth: holdout evidence cannot settle whether a coherent rejected family reflects invalid measurement or a misspecified network [20]. A validity argument chosen after inspecting the same scores it will screen remains circular even if a later split is labeled a holdout (Table 2).

Table 2: Two falsification designs. They answer different questions and are never conflated; the protocol states which design (if any) redefines the headline set.

Design	Admission uses	Holdout asks	Headline consequence
Restriction-stage split ($\mathscr{R} = \mathscr{R}_S \cup \mathscr{R}_H$)	$\mathscr{R}_S$ only; $\mathscr{R}_H$ never votes on admission	Whether survivors satisfy unseen validity implications; verdicts are findings	Selection set $\mathscr{M}_S^*$, unless the protocol says otherwise
Units split (training / hold frames)	Training units only	Whether selected measures comply on held-out units	Robust set $\mathscr{M}_{\text{robust}}^*$; headline range on the robust set when the protocol so defines

Restriction-stage split. Partition the prespecified restrictions into

$$\mathscr{R} = \mathscr{R}_S \cup \mathscr{R}_H, \qquad \mathscr{R}_S \cap \mathscr{R}_H = \emptyset, \tag{8}$$

and admit on $\mathscr{R}_S$ only:

$$\mathscr{M}_S^* = \big\{ m \in \mathscr{M} : \ m \text{ passes every } r \in \mathscr{R}_S \big\}. \tag{9}$$

The held-out restrictions $\mathscr{R}_H$ then ask whether the survivors satisfy validity implications that never voted on admission. Holdout verdicts are scientific *findings* about the selection rule; under a pure stage split they do not redefine the headline set, which is defined by $\mathscr{R}_S$ unless the protocol says otherwise.

Units split. Partition the units into training and hold frames. Admission uses training units only; the hold frame demands compliance of the selected measures. The robust set is

$$\mathscr{M}_{\text{robust}}^* = \mathscr{M}_{\text{select}}^* \cap \big\{ m : \ m \text{ passes the hold-frame screen} \big\}, \tag{10}$$

and, when the protocol so defines, the headline range is the image of β on $\mathscr{M}_{\text{robust}}^*$, with β evaluated on the full pooled sample. Empty robust sets are valid findings.

Comparative claim. Let $L_H(m)$ be a declared loss over the held-out restrictions. The falsifiable prediction is not that every survivor passes every unseen test; it is that construct-guided selection improves unseen validity relative to competing rules:

$$\mathbb{E}\big[L_H(m) \mid m \in \mathscr{M}_S^*\big] \ < \ \mathbb{E}\big[L_H(m) \mid m \in \mathscr{M}_B^*\big], \tag{11}$$

where $\mathscr{M}_B^*$ is selected by a baseline rule—single-anchor fit, best predictive fit, an LLM judge, or random or convenience selection as the floor. Failure is informative: if survivors do no better than random, the network has not demonstrated predictive scientific content. Pre-data discipline fixes the partition, δ, the baseline rule, and the seed before any holdout data are examined; the freeze/run-id contract hash-pins the network, anchors, stage split, and holdout unit list before holdout scoring [23].

3.5 Sampling uncertainty after data-dependent admission

The headline object remains the plug-in range (7): the image of β on the estimated $\mathscr{M}^*$, with no model and no shrinkage. Because $\mathscr{M}^*$ is estimated, the following reporting layer is *mandatory and additive*:

(i) **Uncertainty band.** A units-resampling bootstrap recomputes slacks, $\mathscr{M}^*$, and the endpoints in every replicate (seeded); report per-side quantiles over non-empty replicates. The resampling unit must be declared so that every slack and every $\beta(m)$ is a function of the resampled units; in the country profiles the unit is the country, and in the cell profiles the unit is the demographic cell. Label the object an *uncertainty band*: selection uncertainty on the pooled sample, not a confidence interval and not a holdout-robustness band. The fixed-selection literature takes the admissible set as given [31, 53]; here selection is estimated.

(ii) **Empty-replicate rate.** The share of replicates with empty $\mathscr{M}^*$ is reported explicitly; the band is conditional on non-empty admissible sets. If that share is large, the empty-replicate rate is the primary sampling-fragility summary.

(iii) **Boundary attribution.** Flag measures whose absolute slack margin satisfies $|\text{margin}_m| \leq \kappa \cdot \text{SE}_m$ ($\kappa = 2$ default): near-threshold admissions and near-miss rejections are fragile; far-rejected measures are not.

(iv) **Conservative cross-check (optional).** Compare a units bootstrap that re-forms $\mathscr{M}^*$ in every replicate with the same bootstrap that freezes $\mathscr{M}^*$ at the original estimate; only that contrast isolates admission–endpoint coupling from ordinary endpoint sampling error. A Bonferroni union computed on a fixed $\mathscr{M}^*$ measures interval conservatism, not coupling.

(v) **Validity-argument sensitivity (executed).** The τ-grid varies numeric thresholds while holding the validity theory fixed; sensitivity across substantively defensible networks $\mathscr{R}_1, \mathscr{R}_2, \ldots$ is a separate layer, and Section 5's application executes it: seven networks spanning education floors $\theta \in [0.10, 0.40]$ plus the unrestricted special case are run on the same frozen cell scores (Appendix G). The patience result is invariant over the plausible band—$\mathscr{B}^*(\mathscr{R}) = [0.245, 0.556]$ for $\theta \leq 0.25$, narrowing upward to $[0.417, 0.556]$—while the trust result is argument-dependent: a lenient floor admits a second cheap score whose positive β moves the range across zero, a very strict floor empties the set entirely. Which conclusions are robust to the argument, and which are artifacts of one particular boundary choice, is exactly what this layer exposes.

Under uniform consistency of the slacks and the target functional, and a separation condition excluding measures exactly on the admissibility boundary, the plug-in admissible set and range converge in probability to their population counterparts (Appendix A). The documented failure regime is coupling: a near-boundary measure that is also β-extreme couples admission with the endpoints, because its inclusion or exclusion moves the range directly.

3.6 Scope and composition

The framework covers finite, scalarized menus: every candidate reduces to a declared scalar (or scalarized) object, so survey items, composites, and trained policies enter symmetrically. It is level-agnostic as an interface: a menu exists at every level of the latent hierarchy—item scores, composite scores, population aggregates—and the same screen applies at each level. The evidential meaning is not level-agnostic: single items, composites, and latent scores admit different psychometric properties, and upstream score-construction choices can move both slacks and $\beta(m)$. Composition with causal analysis is not instantiated in this paper and is left as future work. If a design with assumptions $\mathscr{A}$ yields an identified set $\Theta(m; \mathscr{A})$ for each fixed measure, a candidate measurement-robust causal object would be

$$\bigcup_{m \in \mathscr{M}^*} \Theta(m; \mathscr{A}), \tag{12}$$

the union over the admissible set. Construct identification would determine which measures may carry the interpretation; causal identification would proceed within the survivors. Neither step substitutes for the other.

4 An auditable implementation

`cvprofiles` is the open Python implementation of the measurement-selection layer [23]. Its design goal is modest: accept a frozen unit-by-measure score table, a researcher-authored set of restrictions, thresholds, and a downstream functional; return a full measure-by-restriction profile, the admissible set, the robust result set, and the sensitivity diagnostics. The engine implements the state machine

$$\texttt{SCORE} \longrightarrow \texttt{RESTRICT} \longrightarrow \texttt{IDENTIFY} \longrightarrow \texttt{REPORT}, \tag{13}$$

and accepts four plain-text inputs—scores, roles, network, and target specification—producing slacks, $\mathscr{M}^*$, $\mathscr{B}^*$, a survivors-only range, holdout verdicts, and the additive diagnostics of Section 3.5. Frozen runs hash the score matrix, network, and target specification into a reproducibility record, and an empty $\mathscr{M}^*$ is an exit-success scientific finding rather than an exception. The engine is deliberately score-agnostic: it contains no language model, searches no prompt space, and writes no theory, and its validity layer fits no likelihood, prior, or null. Score-agnostic does not mean upstream measurement models are irrelevant—a composite's reliability or an LLM's prompt protocol shapes the score columns and hence the slacks—but the engine itself is agnostic about how scores were produced. This separation is a paper–package contract: the package is released in the public repository (tag `v3.0.2`, version 3.0.2, released on GitHub and PyPI) and used here as an instrument. Version 3.0.1 added a named `empty_R` contract so that $\mathscr{R} = \emptyset$ is expressible

and fail-loud rather than a silent admit-all; default `empty_R=false` is omitted from the network hash, so pre-3.0.1 freezes remain bit-stable. Version 3.0.2 pins round-trip CSV parsing so run identities re-verify from committed inputs on every platform; all runs reported here are 3.0.2 freezes. Software cannot resolve disagreements about what patience or trust means; it can make those disagreements inspectable. Competing researchers can author different networks, thresholds, or menus and see exactly which commitment changes the conclusion—a feature, not a hidden optimizer.

5 Two constructs, two resolutions

The framework is applied to patience and trust at two resolutions, under networks and thresholds frozen before slacks were inspected. Human columns are the GPS scores, WVS Wave 7 items and composites, and a seeded noise control. Two open-weight log-probability columns—Meta-Llama-3.1-8B-Instruct and Phi-4-mini, Q8_0, temperature 0, sha-pinned—are scored from scratch on the published GPS qualitative items and, for trust, the WVS Q57 stem, then reduced to one scalar per model (Appendix E). Country identity is in the prompt; this is silicon sampling, not an unconditioned policy. The two models see identical prompt bytes. The powered surface is the cell resolution ($n = 480$). Country profiles ($n = 41$ and $n = 35$) are the complete but underpowered appendix (Appendix F). Country β is the standardized OLS coefficient of the measure on log GDP per capita with a country-mean education control; cell β is the correlation of the measure with the matching GPS cell mean after country demeaning. Headline admission uses the restriction-stage split of Section 3.4. Every run reports an uncertainty band, an empty-replicate rate, and a τ-grid ($n_{\text{boot}} = 1000$, $\alpha = 0.10$, seed 20260814).

5.1 Constructs, frames, and menus

Patience is time preference as operationalized by the Global Preference Survey, whose module was designed by selecting and weighting survey items for predictive validity against incentivized experimental measures and validated across countries [17, 18]. Under the declared validity argument, a patience measure is expected to rise with human capital and not to collapse into risk preference [16]; the strength of prior evidence behind each expectation conditions how much weight the admission decision can bear. Trust is the belief that other people have good intentions, or that most people can be trusted. GPS trust is a single experimentally validated item; WVS Q57 is the canonical survey item; in-group, out-group, and institution facets are distinct objects of trust [7, 22, 49]. Under the declared argument, a trust measure is expected to co-move with institutional quality and to move opposite income inequality [2, 5]. When GPS trust sits in the menu it is a peer candidate, not a bar.

The country frame is the inner join of GPS, WVS Wave 7, and a pinned WDI snapshot, after a respondent floor of 30 and after masking WVS missing codes -1 through -5 without imputation. Patience is observed for 41 countries; trust, which also requires WGI rule of law and a gini series, for 35. Cells are sex-by-age-band means (six bands from 18–24 through 65+) built separately in GPS individual files and in WVS, then inner-joined, with a minimum of 20 respondents per cell and at least six cells per country: 480 cells in 42 countries. The confirmatory cell estimand country-demeans every column that enters $\mathscr{R}$ or β before SCORE. A pooled (undemeaned) cell run remains on disk as a contrast; it is not the within-country conclusion.

The patience country menu is GPS patience, WVS Q13 (thrift), Q14 (perseverance), their declared composite $z(\text{Q13}) + z(\text{Q14})$, the two open-weight columns, and noise. The trust country menu is GPS trust, Q57, in-group, out-group, institution, the mean of the three multi-item facets (Q57 excluded), the two open-weight columns, and noise. Cell menus drop the GPS column from $\mathscr{M}$ and use it only as the β outcome, so recovery cannot be tautological; they add the same two cheap columns. Education never enters the persona: it is the admission bar.

5.2 Networks

Patience at the country level is screened by $\text{Corr}(m, \text{q275_mean}) \geq 0.20$ (select) and $|\text{Corr}(m, \text{risktaking})| \leq 0.35$ (select), with leftover rank-monotonicity in education at 0.15. The education floor sits well below the published country association of patience with years of schooling [17]; the discriminant bar admits the published patience–risk range of 0.230 [17] to 0.358 [28] and still rejects $r \gtrsim 0.6$. Trust at the country level is screened by $\text{Corr}(m, \text{rule_of_law}) \geq 0.30$ and a negative corr_sign on gini at 0.10, with leftover education at 0.20. There is no corr_min against GPS trust, because that column is in the menu. Cell networks keep only the education implication, now on country-demeaned cell

education, with leftover rank-monotonicity; GDP, gini, and rule of law do not vary within country and are not used. The powered cell screen is therefore a one-gate education floor plus a Spearman leftover on the same auxiliary, not a multi-implication nomological network in the Cronbach–Meehl sense. Multi-restriction screening is exercised at country level (underpowered) and in the nested Campbell–Fiske 2×2.

5.3 Nested special cases

On the present seven-measure country patience menu (GPS, Q13, Q14, composite, Llama, Phi, noise), setting $\mathscr{R} = \emptyset$ in `cvprofiles` 3.0.2 admits the full menu and returns $[L,U] = [-0.219, 0.565]$, the minimum and maximum of those seven β values (endpoints: composite -0.219; Phi 0.565). The composition claim of Section 3.3 is therefore a computed result, not a verbal nesting. An unrestricted multiverse of patience operationalizations still sign-switches on the development estimand. The unrestricted range is reported by design alongside the screened range: the contrast shows how much apparent fragility the validity argument removes and how much survives it. The two objects answer different questions—heterogeneity across plausible operationalizations versus heterogeneity among operationalizations licensed by $\mathscr{R}(C)$—and their disagreement is not a defect. That contrast is the extreme-bounds lesson restated at the measurement layer: Leamer's bounds ranged over specifications with no admissibility test [36], and the "everything is fragile" verdict [48] is precisely what a validity screen prevents—conditional on the validity argument itself being correct (Assumption A.2).

Campbell and Fiske's 2×2 is written as eight ordinary restrictions on a peer menu of GPS patience, WVS child-qualities, GPS trust, and Q57, with $\tau_{\text{conv}} = 0.30$ and $\tau_{\text{disc}} = 0.35$ ($n = 41$). The convergent floor sits at a published survey–behavioral trust bound and below the Falk et al. GPS–WVS trust Spearman of 0.49 [17]; the discriminant ceiling admits their country Corr(patience, trust) of 0.190. Both the classical per-instrument retain set and the engine's global $\mathscr{M}^*$ are empty. The binding cells are not subtle: GPS patience versus child-qualities is $r = -0.252$; GPS trust versus Q57 is $r = 0.284$, just under the floor; Q57 versus GPS *patience* is $r = 0.825$. Partialling log GDP per capita and education leaves that last association at 0.706. The inequalities are Campbell–Fiske cells, so the original correlational MTMM criteria are nested as a special case of the general restriction framework. This recovers the classical correlation-based MTMM logic; modern latent-variable MTMM (e.g., trait–method models such as CT-C($M-1$)) is a broader family this framework does not implement. Substantively the 2×2 rejects every instrument: at country level the canonical WVS trust item is closer to GPS time preference than to GPS "best intentions." That is a finding about the named special case, not a reason to lower τ after seeing slacks.

5.4 Within-country cells

Table 3 is the powered application. Country-demeaning the 480 cells is required: Q13 was rejected in the undemeaned table because richer countries have both more education and lower thrift endorsement. One competing explanation for such cross-resolution reversals is measurement noninvariance: if a multi-item measure's structure differs across the pooled versus within-country levels, the same nominal column need not represent the same construct at both, so part of a reversal can reflect a changed measurement function rather than a changed construct relation. The diagnostic of Appendix J separates the two mechanisms for this application: thrift's reversal is level-dependence of the education *expectation* (its sign flips inside countries), not item-level differential functioning; perseverance is uniformly negative at every level. Invariance and differential-item-functioning evidence from item-level microdata remains the stronger test and a natural extension [57] (Section 6.7). Inside countries the sign flips. Q13's education slack is $+0.051$, leftover monotonicity $+0.060$, and β versus GPS patience $+0.245$. Q14 stays anti-aligned (slack -0.564, $\beta = -0.326$). The composite is Q13 dragged under the bar by Q14. The rejected family is coherent: Q13, Q14, and their composite all point negatively on the development estimand at country level, and Q14 remains anti-aligned inside countries. Under the reciprocal reading of construct validation, that coherence is evidence—of measurement failure or of a divisible construct—and a candidate input to theory revision, not merely noise the screen removed.

Both open-weight patience columns clear the same education bar with room (slacks $+0.322$ and $+0.440$) and leftover monotonicity. Their cell βs are $+0.417$ and $+0.556$. The admissible set is $\{\text{Q13}, \text{Llama}, \text{Phi}\}$ with plug-in range $[0.245, 0.556]$, cell-exchangeable band $[0.179, 0.614]$, and an empty-replicate rate of zero. Without the cheap columns the set was the Q13 singleton 0.245 with empty-replicate rate 0.17. Q13's education slack remains knife-edge; $\lambda = 1.5$ still empties a human-only menu. The cheap columns are not knife-edge on that bar. They do not make Q14 admissible, and they are not a second method: they share architecture, pretraining-era geographic associations, and the

Table 3: Powered application: country-demeaned cells with fresh open-weight columns (`cvprofiles` 3.0.2, seed 20260814, $n_{\text{boot}} = 1000$). Cell β is the correlation with the matching GPS cell mean after country demeaning. The cell-exchangeable band is an additive uncertainty band at $\alpha = 0.10$, not a confidence interval. The country-cluster band resamples countries with replacement and re-demeans within each replicate (cluster unit is country). Empty-rep. is the cell-exchangeable rate; cluster empty-rep. is 0.000 for patience and 0.001 for trust. Country profiles are Appendix F.

Profile	n	$\mathscr{M}^*$	$[L,U]$	**Cell-exch. band**	**Cluster band**	**Empty-rep.**
Patience, cells	480	{Q13, Llama, Phi}	$[0.245, 0.556]$	$[0.179, 0.614]$	$[0.174, 0.622]$	0.000
Trust, cells	480	{Phi}	-0.317	$[-0.382, -0.240]$	$[-0.400, 0.195]$	0.000

same prompts. The patience range widens from the Q13 singleton 0.245 to $[0.245, 0.556]$ because both cheap columns clear one education bar under identical prompts—menu expansion under correlated generators, which Section 1.2 already flags as weak validity evidence.

Every WVS trust facet still fails the education bar after demeaning, including at $\lambda = 0.5$. Llama fails it as well (slack -0.077). Phi passes (slack $+0.149$, leftover $+0.133$) and is the singleton. Its plug-in β versus GPS trust is -0.317, with a cell-exchangeable band $[-0.382, -0.240]$ entirely negative; a country-cluster band on that singleton crosses zero (Table 3). GPS trust itself fails rule of law at country level ($r = 0.12$ versus floor 0.30), so the cell β is criterion-recovery of that contested column, not a refutation of a validated criterion. In-group, out-group, institution, and the composite still co-move with GPS trust (0.224, 0.211, 0.285, 0.318) without tracking within-country education; Q57 does not (0.052). The methods result is Appendix H Finding 3: among human trust-cell measures, education slack and β are anti-aligned (Spearman/Pearson ≈ -0.80).

5.5 What the application is allowed to mean

Validity decides which measures may enter the conditional robustness set; rejected measures are not epistemically silent, and a coherent rejected family is evidence about the adequacy of $\mathscr{R}(C)$ and of the construct definition—the input to a discovery channel (revise, refreeze, test on new data) that this paper does not execute. The range reports leftover disagreement among those admitted; the band reports that the admissible set itself is noisy. The inadmissible set is profiled directly in Appendix H: rejection is restriction-concentrated rather than scattered, and the alignment between failing the screen and disagreeing with the estimand runs in opposite directions for the two constructs—perfectly positive for patience, strongly negative for trust.

Read that way, the cell application does not deliver a preferred measure of patience or of trust. Thrift and perseverance are not interchangeable proxies. Two open-weight patience scores survive the same bar as thrift and pull GPS recovery from a knife-edge 0.245 to $[0.245, 0.556]$. The same education implication empties the human trust menu and admits only Phi; the plug-in association with GPS trust is $\beta = -0.317$, recovery of a contested criterion. Llama and Phi do not agree on trust at either resolution (Appendix F). Those are disagreements the composition is designed to make inspectable. They are not a license to move τ, to treat two GGUFs as two methods, or to call the menus a measure of culture.

Country samples are small, so the country profiles—including the cheap columns—are the complete underpowered appendix rather than the headline. Thresholds remain researcher-owned. Cell country-demeaning was a dated estimand amendment, not a post-hoc rescue of θ. Circularity is a standing risk rather than a solved problem.

Table 4 reports the cell diagnostics already computed in the coverage artifacts. Table 5 compares construct-guided selection with best-β, best-anchor, and size-matched random subsets on the frozen cell slacks.

6 What follows—and what does not

6.1 Transparency and openness

Data and materials. Frozen score matrices, roles, networks, thresholds, and all run artifacts behind the reported tables are committed in the public repository (`https://github.com/Bonorinoa/cvprofiles`, tag `v3.0.2`). Raw WVS Wave 7 and GPS microdata are license-restricted and are not redistributed; the derived country- and cell-level score tables are committed. Open-weight model artifacts (Meta-Llama-3.1-8B-Instruct and Phi-4-mini, Q8_0) are pinned by

Table 4: Cell-profile diagnostics from the frozen coverage artifacts ($\kappa = 2$; margin is the binding slack). Load-bearing: Q13 is a knife-edge admission ($\hat{p} = 0.83$); Llama-trust is a near-miss rejection ($\hat{p} = 0.022$). Cell-exchangeable frozen-M^* versus re-formed bands are nearly the same width; coupling is $\hat{p}$(Q13), not extra band width.

Measure	In $\mathscr{M}^*$	β	Sel. slack	Hold. slack	Margin	SE	κ=2	$\hat{p}$
Patience cells								
Q13	yes	0.245	0.051	0.06	0.051	0.051	yes	0.83
Q14	no	−0.326	−0.564	−0.499	−0.564	0.044	no	0.0
composite	no	−0.027	−0.25	−0.222	−0.25	0.058	no	0.0
Llama	yes	0.417	0.322	0.358	0.322	0.034	no	1.0
Phi	yes	0.556	0.44	0.523	0.44	0.029	no	1.0
noise	no	−0.003	−0.231	−0.165	−0.231	0.046	no	0.0
Trust cells								
Q57	no	0.052	−0.177	−0.131	−0.177	0.049	no	0.0
in-group	no	0.224	−0.268	−0.233	−0.268	0.043	no	0.0
out-group	no	0.211	−0.275	−0.225	−0.275	0.051	no	0.0
institution	no	0.285	−0.627	−0.552	−0.627	0.04	no	0.0
composite	no	0.318	−0.474	−0.414	−0.474	0.046	no	0.0
Llama	no	0.179	−0.077	−0.066	−0.077	0.04	yes	0.022
Phi	yes	−0.317	0.149	0.133	0.133	0.043	no	1.0
noise	no	0.015	−0.172	−0.119	−0.172	0.048	no	0.0

Table 5: Comparative falsification on frozen cell slacks. $H(m) = 1$ iff every leftover holdout slack ≥ 0. Random is 1000 size-k subsets. At cell resolution the leftover is a rank transform of the selecting education correlation, so the patience comparison is weak; the trust comparison is not.

Profile	$\mathscr{M}_S^*$	$\hat{p}_S$	Best-β	Best-anchor	Rand. mean	Share $\geq \hat{p}_S$
Patience	{Q13, Llama, Phi} (k=3)	1.0	Phi ($\hat{p}$=1.0)	Phi ($\hat{p}$=1.0)	0.49	0.053
Trust	{Phi} (k=1)	1.0	composite ($\hat{p}$=0.0)	Phi ($\hat{p}$=1.0)	0.129	0.129

SHA-256 checksum, and the scored columns are committed. A versioned deposit of the run artifacts with a persistent DOI will be made at acceptance.

Code. The engine is the open-source package `cvprofiles` (version 3.0.2, MIT license), with a documented state machine (SCORE → RESTRICT → IDENTIFY → REPORT), frozen-run hash contracts, and an executable regression test suite; the same release is on PyPI.

Preregistration. This study was not preregistered in a public registry. A freeze is not prospectivity: it is auditable conditional on a stated moment. Networks and thresholds were authored and hash-pinned before slacks were inspected, the restriction-stage split of Section 3.4 holds leftover restrictions that never vote on admission, and the freeze/run-id contract pins scores, network, and target at that stated moment.

Reporting of AI use. Language models were used as measurement instruments: the open-weight columns (Appendix E) were produced by prompted, temperature-zero scoring of published survey items under a documented adapter card. An AI writing assistant (Hermes Agent, Nous Research [45]) was used for editing and formatting assistance; all scientific content, analysis, and research decisions are the authors'. No AI system is an author.

6.2 What an empty or wide set means

An empty $\mathscr{M}^*$ rejects the pair consisting of the current menu and the validity argument on the stated sampling frame; it does not refute the existence of the construct. Conditional on $\mathscr{R}$, a wide $\mathscr{B}^*(\mathscr{R})$ says the conclusion remains sensitive across the surviving operationalizations, and a narrow $\mathscr{B}^*(\mathscr{R})$ says it is homogeneous among them; the unrestricted image over $\mathscr{M}$ answers a different question—heterogeneity across plausible operationalizations—and is reported alongside, not instead. All of these outcomes are scientifically interpretable, and none should be tuned away by relaxing δ after seeing the result.

6.3 The hierarchy of uncertainties

The paper does not claim to discover a new category called measurement uncertainty. Psychometrics, metrology, latent-variable models, and multiverse work have long separated measurement from sampling and analytic choices. The claim is compositional: construct validity determines which measurement alternatives deserve scientific standing; measurement robustness propagates the unresolved choice among them; specification robustness can then vary analytic decisions conditional on each admissible measure; and sampling inference quantifies repeated-sample uncertainty within that design. The layers answer different questions and should not be collapsed into one standard error. The relationship between the layers is also reciprocal rather than one-way: the validity argument adjudicates admission for conditional inference, and structured disagreement in the full menu is admissible evidence against the argument in the next iteration.

6.4 Beyond social-science measures

Benchmark scores are measurement functions too: they map model behavior into quantities interpreted as reasoning, safety, or knowledge. A family of task families can therefore be treated as a menu, with held-out task families and cross-method evidence supplying a validity network (Appendix C). The same logic applies to medical risk scores, remote-sensing proxies, and administrative indices. Domain-specific validity evidence changes; the composition does not.

6.5 Researcher-owned validation gates in automated science

Automation can expand the menu, compute scores, and execute a frozen network. It should not silently redefine the construct, rewrite the network after seeing results, or treat correlated model outputs as independent evidence. A credible automated-science architecture therefore needs a researcher-owned, contestable validation gate—one that can itself fail, and whose commitments can be compared across research teams (Appendix D). The most direct failure mode is Goodharting: a generator trained explicitly against the known restrictions. The defenses are procedural rather than magical—freeze the network and thresholds before final scoring, preserve negative controls, reserve restrictions or units, and keep generator provenance separate from validity evidence. Anti-circularity must not become anti-learning: revising a construct or network after structured anomalies is legitimate theory development, provided the revised contract is frozen and tested on new data; what is forbidden is silently rewriting $\mathscr{R}$ or τ after seeing slacks to rescue the current result.

6.6 What this application decides—and what remains

The application is a stress test of the reporting rule. It does not decide what patience or trust *is*. On the powered cell surface, the screen discriminates among semantically related proxies and among cheap generators: thrift and both open-weight patience scores survive and recover GPS patience in $[0.245, 0.556]$, perseverance is anti-aligned, and the only cheap trust score that clears education is negatively associated with GPS trust (recovery of a contested criterion). The unrestricted country patience menu still sign-switches; once the same cheap columns face the country network they join GPS and widen the development range, while the two models disagree on trust (Appendix F).

What remains includes individual-level gradient tests that keep residual variance the cell means discard; open libraries of networks and thresholds so that research teams can contest measurement definitions without silently rewriting them; a Monte Carlo of the band; prompt-paraphrase of the LLM columns; and a multi-implication cell network on *new* data.

6.7 Open implementation, precision layer, and community

The registry of restriction types, target functionals, and generators is deliberately small, and its gaps are known: stability is schema-only, slacks are never learned, and coupling theory is deferred to a companion methods paper. One gap has a named plan: a *measurement-precision layer*. Admissibility today is judged on reliability-naive sample moments, so two composites with identical latent validity can land on opposite sides of a gate purely because of attenuation (DGP F, Appendix B). The planned extension accepts optional per-measure reliability metadata and reports disattenuated slack and $\beta(m)$ variants alongside the naive ones, so that admissibility decisions can be audited against

the score-construction machinery, not only the substantive construct. The contribution we invite is concrete—new restriction types, β-functionals, and measurement generators—not a leaderboard.

7 Conclusion

Science does not become easier merely because candidate measures become cheap. Abundance moves the bottleneck from producing an operationalization to justifying one. The appropriate response is not to rank human surveys above AI, or fine-tuning above prompting, but to make competing measures answer to the same construct-level evidence. The framework asks six questions in order: What do we mean? How might we measure it? What should a credible measure do? Which measures survive? Which scientific conclusions survive with them? And what does disagreement among the measures teach us about the construct and the argument? Formally,

$$C \longrightarrow \mathscr{M} \xrightarrow{\mathscr{R}} \mathscr{M}^*(\mathscr{R}) \longrightarrow \mathscr{B}^*(\mathscr{R}), \qquad \mathscr{M},\, \mathscr{B}^* \dashrightarrow \text{revision of } C \text{ or } \mathscr{R}.$$

The novelty is not construct validity, multiverse analysis, partial identification, or sampling inference in isolation; it is their operational composition at the measurement layer. When measures are cheap, validity becomes the scarce resource.

Data and code availability

Frozen score matrices, roles, networks, thresholds, and all run artifacts behind the tables are committed in the public repository `https://github.com/Bonorinoa/cvprofiles` at tag `v3.0.2`; version 3.0.2 is on PyPI. Raw WVS Wave 7 and GPS microdata are license-restricted and are not redistributed; the derived country and cell score tables are committed. Open-weight model artifacts (Meta-Llama-3.1-8B-Instruct and Phi-4-mini, Q8_0) are pinned by SHA-256 checksum, and the scored columns are committed. A versioned DOI will be deposited on Zenodo at acceptance.

Author contributions

Augusto Gonzalez-Bonorino: conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft, visualization. Daniel J. Martin: methodology, writing—review and editing. (PLACEHOLDER)

Author note

This work has not been presented at a conference or posted as a preprint. The companion software package `cvprofiles` is publicly released (GitHub tag v3.0.2; PyPI version 3.0.2). Generative AI was used in two ways: open-weight language models were used as measurement instruments (documented in Appendix E), and an AI writing assistant (Hermes Agent, Nous Research) was used for editing and formatting assistance; all scientific content, analyses, decisions, and nomological networks are the authors'.

References

[1] Agarwal, D., Shukla, A., Goyal, T., & Vashistha, A. (2026). *PLURAL: A global dataset for value alignment* [Preprint]. arXiv. https://arxiv.org/abs/2607.08034

[2] Aghion, P., Algan, Y., Cahuc, P., & Shleifer, A. (2010). Regulation and distrust. *Quarterly Journal of Economics, 125*(3), 1015–1049. https://doi.org/10.1162/qjec.2010.125.3.1015

[3] Baumann, J., Röttger, P., Urman, A., Wendsjö, A., Plaza-del-Arco, F. M., Gruber, J. B., & Hovy, D. (2025). *Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation* [Preprint]. arXiv. https://arxiv.org/abs/2509.08825

[4] Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Foroutan, N., Schmitz, C., Korgul, K., Batra, H., Deb, O., Beharry, E., Emde, C., Foster, T., Gausen, A., Grandury, M., Han, S., Hofmann, V., Ibrahim, L., … Mahdi, A. (2025). *Measuring what matters: Construct validity in large language model benchmarks* [Conference paper]. NeurIPS 2025 Datasets and Benchmarks Track. https://arxiv.org/abs/2511.04703

[5] Bjørnskov, C. (2008). Social capital and happiness in the United States. *Social Indicators Research, 86*(1), 269–282. https://doi.org/10.1007/s11205-007-9106-8

[6] Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

[7] Butler, J. V., Giuliano, P., & Guiso, L. (2016). The right amount of trust. *Journal of the European Economic Association, 14*(5), 1155–1188. https://doi.org/10.1111/jeea.12178

[8] Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

[9] Canen, N., & Enamorado, T. (2026). *When predictions become regressors: A split-sample correction for biases in downstream inference* [Preprint]. arXiv. https://arxiv.org/abs/2608.02909

[10] Chen, X., Rambachan, A., & Tamer, E. (2026). *Partial identification from LLM prompts* [Preprint]. arXiv. https://arxiv.org/abs/2606.15031

[11] Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

[12] Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920954925

[13] Dell, M., & Rambachan, A. (2026). *Estimation and inference with AI-generated data* [Paper presentation]. NBER Summer Institute 2026 Methods Lectures.

[14] Desai, M., Card, D., & Jacobs, A. Z. (2026). *Validating LLMs in social science: Epistemic threats and emerging norms* [Preprint]. arXiv. https://arxiv.org/abs/2607.07915

[15] Dey, P., Joshi, B., Poddar, P., Zhao, J., & Ferrara, E. (2026). *PALMs: Using multi construct-grounded rationales for modeling population preferences in LLMs* [Preprint]. arXiv. https://arxiv.org/abs/2608.01458

[16] Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association, 9*(3), 522–550. https://doi.org/10.1111/j.1542-4774.2011.01015.x

[17] Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *Quarterly Journal of Economics, 133*(4), 1645–1692. https://doi.org/10.1093/qje/qjy013

[18] Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science, 69*(4), 1935–1950. https://doi.org/10.1287/mnsc.2022.4455

[19] Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science, 362*(6412), Article eaas9899. https://doi.org/10.1126/science.aas9899

[20] Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

[21] Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

[22] Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *Quarterly Journal of Economics, 115*(3), 811–846. https://doi.org/10.1162/003355300554926

[23] Gonzalez-Bonorino, A. (2026). *cvprofiles: Construct-validity profiles as partial identification over measurement menus* (Version 3.0.2) [Computer software]. GitHub. https://github.com/Bonorinoa/cvprofiles

[24] Gonzalez-Bonorino, A., Capra, C. M., & Pantoja, E. (in press). LLMs can model non-WEIRD behavior: Experiments with synthetic cultural agents. In *Artificial intelligence and experimental methods* (Research in Experimental Economics, Vol. 23). Emerald. https://arxiv.org/abs/2501.06834

[25] Hall, A. B. (2026). *AI research topics at PolMeth 2026* [Data set]. GitHub. https://github.com/andybhall/polmeth-2026-ai

[26] Halterman, A., & Keith, K. A. (2026). Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis, 34*(2), 188–204. https://doi.org/10.1017/pan.2025.10017

[27] Hanel, P. H. P., & Zarzeczna, N. (2023). From multiverse analysis to multiverse operationalisations: 262,143 ways of measuring well-being. *Religion, Brain & Behavior, 13*(3), 309–313. https://doi.org/10.1080/2153599X.2022.2070259

[28] Hanushek, E. A., Kinne, L., Lergetporer, P., & Woessmann, L. (2022). Patience, risk-taking, and human capital investment across countries. *The Economic Journal, 132*(646). https://doi.org/10.1093/ej/ueab105

[29] Horton, J. J., Filippas, A., & Manning, B. S. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? (NBER Working Paper No. 31122). National Bureau of Economic Research. https://doi.org/10.3386/w31122

[30] Hullman, J., Broska, D., Sun, H., & Shaw, A. (2026). *This human study did not involve human subjects: Validating LLM simulations as behavioral evidence* [Preprint]. arXiv. https://arxiv.org/abs/2602.15785

[31] Imbens, G. W., & Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica, 72*(6), 1845–1857. https://doi.org/10.1111/j.1468-0262.2004.00555.x

[32] Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

[33] Kiet, H. T., Minh, D. S. D., Nguyen, T., Tran, C.-N., Pham, P.-H., Quy, N. L. P., Han, T. A., & Tran-Thanh, L. (2026). *Training-free cultural alignment of large language models via persona disagreement* [Preprint]. arXiv. https://arxiv.org/abs/2605.10843

[34] Kim, E. M., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning* (PMLR 267, pp. 30038–30066).

[35] Kirgis, P., Kapoor, S., Schwartz, A., Rabanser, S., Africa, D., Voudouris, K., Nguyen, V., Pilditch, T., Dubois, M., Coppock, H., Ududec, C., Nadgir, N., Orona, M., Bayer, T., Chan-Sew, D., Ling, Y., Shetty, A., Toner, H., Hadfield, G., … Narayanan, A. (2026). *Can AI agents conduct open-ended AI research? Early evidence from two case studies* [Preprint]. arXiv. https://arxiv.org/abs/2607.27191

[36] Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review, 73*(1), 31–43.

[37] Li, B., Yu, T., Koa, K. J. L., & Huang, K.-W. (2026). The proxy presumption: From semantic embeddings to valid social measures. In *Proceedings of ACL 2026* (pp. 22892–22910). https://doi.org/10.18653/v1/2026.acl-long.1048

[38] Licht, H., Sarkar, R., Wu, P. Y., Goel, P., Stoehr, N., Ash, E., & Hoyle, A. M. (2025). *Measuring scalar constructs in social science with LLMs* [Conference paper]. EMNLP 2025. https://arxiv.org/abs/2509.03116

[39] Lin, V., Yun, T., Matarić, M., Canny, J., Gretton, A., & D'Amour, A. (2026). *The illusion of intervention: Your LLM-simulated experiment is an observational study* [Preprint]. arXiv. https://arxiv.org/abs/2605.20767

[40] Manning, B. S., Zhu, K., & Horton, J. J. (2024). *Automated social science: Language models as scientist and subjects* (NBER Working Paper No. 32381). National Bureau of Economic Research. https://doi.org/10.3386/w32381

[41] Manski, C. F. (2003). *Partial identification of probability distributions*. Springer.

[42] Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

[43] Messing, S. (2026). *Hidden measurement error in LLM pipelines distorts annotation, evaluation, and benchmarking* [Preprint]. arXiv. https://arxiv.org/abs/2604.11581

[44] Molinari, F. (2020). Microeconometrics with partial identification. In *Handbook of econometrics* (Vol. 7A, pp. 355–486). Elsevier.

[45] Nous Research. (2026). *Hermes Agent* [Computer software]. https://hermes-agent.nousresearch.com

[46] Oğut, B., & Yin, M. (2026). *Partial identification with multiple nonlinear measurements of a latent regressor* [Preprint]. arXiv. https://arxiv.org/abs/2607.12219

[47] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems* (Vol. 36).

[48] Sala-i-Martin, X. (1997). I just ran two million regressions. *American Economic Review, 87*(2), 178–183.

[49] Sapienza, P., Toldrá-Simats, A., & Zingales, L. (2013). Understanding trust. *The Economic Journal, 123*(573), 1313–1332. https://doi.org/10.1111/ecoj.12036

[50] Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour, 4*, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

[51] Stanford Institute for Human-Centered Artificial Intelligence. (2026). *AI index report 2026* [Report]. Stanford University. https://hai.stanford.edu/ai-index/2026-ai-index-report

[52] Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

[53] Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica, 77*(4), 1299–1315. https://doi.org/10.3982/ECTA7347

[54] Truong, S., Wang, S., Wang, A., Truong, N., Reuel, A., Haber, N., & Koyejo, S. (2026). *Public AI benchmarks are broken, but are private benchmarks the answer?* [Manuscript in preparation]. https://aimslab.stanford.edu/papers/Truong2026PublicPrivate.pdf

[55] Wallach, H., Desai, M., Pangakis, N., Cooper, A. F., Wang, A., Barocas, S., Chouldechova, A., Atalla, C., Blodgett, S. L., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Reed, S., Sheng, E., Vann, D., Vaughan, J. W., Vogel, M., Washington, H., & Jacobs, A. Z. (2024). Evaluating generative AI systems is a social science measurement challenge. In *NeurIPS 2024 Workshop on Evaluating Evaluations (EvalEval)*. https://arxiv.org/abs/2411.10939

[56] Yin, M., Vu, H., & Persico, C. (2026). *How (un)stable are LLM occupational exposure scores? Evidence from multi-model replication* (NBER Working Paper No. 35110). National Bureau of Economic Research. https://doi.org/10.3386/w35110

[57] Zeinfeld, L., Strugatski, A., Bar-Dov, Z., & Blonder, R. (2026). *Assessment design in the AI era: A method for identifying items functioning differentially for humans and chatbots* [Preprint]. arXiv. https://arxiv.org/abs/2603.23682

A First-principles formalization

A.1 Primitives

Definition A.1 (Measurement function and menu)**.** *A* measurement function *is a map $m : \mathcal{X} \to \mathbb{R}$ (or to a finite choice simplex) producing a score or response distribution from available inputs $X \in \mathcal{X}$; a map to a vector followed by a* declared *scalar reduction is part of the measure's definition. A* menu *is a finite collection $\mathcal{M} = \{m_1, \ldots, m_J\}$ declared by the researcher.*

Definition A.2 (Nomological restriction)**.** *Given auxiliaries V and a threshold $\tau_r \in \mathbb{R}$, a* restriction *is a functional inequality $\mathbb{E}[g_r(m(X), V; \tau_r)] \geq 0$. The* sample slack $s_r(m)$ *is a sample analogue of the left-hand side; the sign convention is $s_r(m) \geq 0$ when the restriction is satisfied.*

Definition A.3 (Admissible set and admissible-result set)**.** *For tolerance $\delta \geq 0$,*

$$\mathcal{M}^* = \{m \in \mathcal{M} : \; s_r(m) \geq -\delta \; \forall r \in \mathcal{R}\}.$$

For a target functional $\beta : \mathcal{M} \to \mathbb{R}$,

$$\mathcal{B}^* = \{\beta(m) : m \in \mathcal{M}^*\}, \qquad [L, U] = [\min \mathcal{B}^*, \; \max \mathcal{B}^*] \quad (\mathcal{M}^* \neq \emptyset).$$

Definition A.4 (Menu-level validity)**.** *Inference is* menu-level *when the object of scientific interest is $\mathcal{M}^*$ (and functionals of $\mathcal{M}^*$), not a single preferred m. Per-measure protocols [37, 26] validate individual operationalizations; they are complementary inputs to a menu, not substitutes for menu-level inference.*

A.2 Population versus sample

The population slack of restriction r on measure m is $s_r^0(m) = \mathbb{E}[g_r(m(X), V; \tau_r)]$. The population admissible set and image are

$$\mathcal{M}_0^* = \{m \in \mathcal{M} : s_r^0(m) \geq -\delta \; \forall r \in \mathcal{R}\}, \qquad \mathcal{B}_0^* = \{\beta(m) : m \in \mathcal{M}_0^*\}.$$

Hats $\hat{s}_r(m)$, $\hat{\mathcal{M}}^*$, $\hat{\mathcal{B}}^*$, $[\hat{L}, \hat{U}]$ are the sample analogues computed from the frozen score matrix. Section 3.3 uses s_r as that sample slack when it defines $\mathcal{M}^*$; $\mathcal{M}^*$ is not a partial-identification set for a latent construct. $\mathcal{M}^*$ is observed-score suitability for a declared use (DGP F); a measurement-precision layer is unimplemented.

A.3 Maintained assumptions

Assumption A.1 (Menu coverage)**.** *$\mathcal{M}$ contains the operationalizations of scientific interest for the research question; coverage is argued from literature and design, never proven.*

Assumption A.2 (Restriction validity)**.** *Each g_r is a correct testable implication of the construct definition and auxiliaries V; restrictions are researcher-owned and falsifiable by data.*

Assumption A.3 (Sampling)**.** *The unit-by-measure score matrix is drawn under a stated sampling model (e.g., exchangeable country-level aggregates), frozen with the run.*

Assumption A.4 (Tolerance policy)**.** *$\delta \geq 0$ is declared ex ante (default 0) and is never tuned to rescue an empty set. A common additive tolerance $\delta > 0$ is not scale-invariant across restrictions; this application uses $\delta = 0$.*

A.4 Empty restriction sets and empty admissible sets

Proposition A.1 (Empty restriction set and empty admissible set)**.** *If $\mathcal{R} = \emptyset$, there is no restriction index over which to take a maximum: every $m \in \mathcal{M}$ is admissible, $\mathcal{M}^* = \mathcal{M}$, and $[L, U]$ is the unrestricted image over the declared menu. If $\mathcal{M}^* = \emptyset$, the range $[L, U]$ is undefined.*

A.5 Consistency of the plug-in objects

Proposition A.2 (Consistency of the admissible set and range)**.** *Assume A1–A4. Suppose*

$$\max_{m\in\mathscr{M}}\max_{r\in\mathscr{R}}\left|\hat{s}_r(m)-s_r(m)\right| \xrightarrow{p} 0, \qquad \max_{m\in\mathscr{M}}\left|\hat{\beta}(m)-\beta(m)\right| \xrightarrow{p} 0,$$

and suppose separation: the slack margin $\mu(m) := \min_r s_r(m) + \delta$ *satisfies* $\mu(m) \neq 0$ *for all* $m \in \mathscr{M}$*. Then* $\Pr(\hat{\mathscr{M}}^* = \mathscr{M}^*) \to 1$ *and* $[\hat{L},\hat{U}] \xrightarrow{p} [L,U]$.

Proof sketch. Let $\eta := \min_m |\mu(m)| > 0$ by separation. Uniform consistency gives $\Pr(\max_{m,r}|\hat{s}_r - s_r| < \eta/2) \to 1$. On that event, $\hat{\mu}(m)$ has the sign of $\mu(m)$ for every m, so $\hat{\mathscr{M}}^* = \mathscr{M}^*$; on the same event the endpoints are extrema of $\hat{\beta}$ over the fixed finite set $\mathscr{M}^*$, and the asserted convergence follows from the uniform bound on $\hat{\beta}$. Boundary measures with $\mu(m) = 0$ are the only obstruction: their admission is decided by sampling noise. □

Remark A.1 (Empty sets)**.** $\mathscr{M}^* = \emptyset$ *rejects the pair* $(\mathscr{M},\mathscr{R})$*, not the construct C; it is a finding, not an error.*

Remark A.2 (Consistency is not validity)**.** *Proposition A.2 says the plug-in objects are well-posed. Validity claims come from the restrictions themselves (A2) and from the informativeness of* $\mathscr{B}^*$.

A.6 Holdout null and claim

Assumption A.5 (Pre-data discipline)**.** *The partition* $(\mathscr{R}_S,\mathscr{R}_H)$*, tolerance* δ*, baseline rule B, and random seed are fixed before any holdout data are examined.*

Proposition A.3 (Finite-menu pass proportion)**.** *Under pre-data discipline,* $\mathscr{M}_S^*$ *does not depend on* $\mathscr{R}_H$*. Let* $H(m) = \mathbf{1}\{m \text{ passes every } r \in \mathscr{R}_H\}$*. On a finite declared menu the reported object is the pass proportion*

$$\hat{p}_S = |\mathscr{M}_S^*|^{-1} \sum_{m\in\mathscr{M}_S^*} H(m)$$

when $\mathscr{M}_S^* \neq \emptyset$*, compared with the analogue on a baseline selected set* $\mathscr{M}_B^*$*; it is not an expectation over measures without a probability law on* $\mathscr{M}$*. If* $\mathscr{M}_S^* = \emptyset$ *or* $\mathscr{M}_B^* = \emptyset$*, the corresponding pass proportion is undefined.*

A.7 Coupling and the uncertainty band

The plug-in range inherits sampling error from both $\hat{\beta}(m)$ and $\hat{\mathscr{M}}^*$. When a measure is near the boundary and β-extreme, the two errors are coupled. Endpoint inference for fixed identified sets [31, 53] takes selection as given; here selection is estimated. The main-text protocol—uncertainty band, empty-replicate rate, boundary attribution, optional conservative projection—is an uncertainty summary under a stated reading, not a simultaneous confidence region under arbitrary dependence. Sample splitting decorrelates admission and endpoints at the cost of effective sample size. Formal simultaneous-coverage theory is the subject of the companion methods paper.

A.8 Notation at a glance

Symbol	Meaning
C	Scientific construct
m	Candidate measure / measurement function
$\mathscr{M}$	Declared finite menu
V	Auxiliaries used in validity restrictions
$\mathscr{R}(C)$	Validity restrictions / nomological network
τ_r	Substantive threshold (researcher-owned)
$\mathscr{R}_S, \mathscr{R}_H$	Selection and holdout partitions
$s_r^0(m)$	Population slack of restriction r on measure m
$s_r(m)$	Sample slack of restriction r on measure m
δ	Slack tolerance (default 0)
$\mathscr{M}_0^*, \mathscr{B}_0^*$	Population admissible set and image
$\mathscr{M}^*$	Admissible set (sample)
$\mathscr{M}^*_{\text{select}}, \mathscr{M}^*_{\text{robust}}$	Select and robust sets under holdout designs
$\beta(m)$	Scientific result using measure m
$\mathscr{B}^*$	Admissible-result set (construct-identified)
$[L,U]$	Assumption-indexed admissible-measure range
$G(D,C)$	Measurement-generation technology

B Synthetic data-generating processes

Synthetic examples remain useful because population truth is known. Consider standardized latent C, a convergent criterion V_c with $\text{Corr}(V_c, C) = 0.80$, a discriminant foil V_d orthogonal to C, an outcome Y with $\text{Corr}(Y, C) = 0.60$, and a known-groups latent gap of 0.80. The menu contains a high-quality measure loading 0.90 on C, a weak measure loading 0.50, a contaminated measure loading 0.70 on C and 0.55 on V_d, a method-variance-dominated measure loading 0.20, and pure noise (Table 6).

Table 6: Population menu for hand-verifiable synthetic examples.

Measure	Loading on C	Loading on V_d	Method variance	Intended role
m_{true}	0.90	0	residual	High-quality operationalization
m_{weak}	0.50	0	residual	Valid but noisy
m_{contam}	0.70	0.55	residual	Contaminated by discriminant foil
m_{method}	0.20	0	dominant	Method-variance dominated
m_{noise}	0	0	1	Negative control

DGP A: clean separation. Selection restrictions are $\text{Corr}(m, V_c) \geq 0.39$, $|\text{Corr}(m, V_d)| \leq 0.25$, and a known-groups gap of at least 0.25; target $\beta(m) = \text{Corr}(m, Y)$. Population slacks are $+0.33$, $+0.01$, $+0.17$, -0.23, -0.39 on convergence; $+0.25$, $+0.25$, -0.30, $+0.25$, $+0.25$ on discrimination; and $+0.47$, $+0.15$, $+0.31$, -0.09, -0.25 on the group gap. Only m_{true} and m_{weak} survive, and

$$\mathscr{B}^* = \{0.54, 0.30\}, \qquad [L, U] = [0.30, 0.54].$$

Every admissible slack is strictly positive: separation holds.

DGP B: boundary measures and coupling. Raise the convergent threshold to 0.70. Then m_{true} remains admissible by a $+0.02$ margin and m_{weak} is rejected (-0.30); the plug-in range collapses to the singleton $\{0.54\}$. If instead the threshold is set at the weak measure's population correlation (0.40), m_{weak} lies exactly on the boundary: in finite samples its admission is decided by sampling noise, and because it attains the lower endpoint whenever admitted, admission error and endpoint error are coupled—the regime the boundary-attribution protocol monitors.

DGP C: empty $\mathscr{M}^*$ is a finding. Replace the discriminant restriction with $\text{Corr}(m, V_d) \geq 0.50$ and raise the convergent threshold to 0.60. No measure can load heavily on both C (hence on V_c) and the orthogonal foil at once: m_{true} passes convergence ($+0.12$) but fails the foil (-0.50); m_{contam} passes the foil ($+0.05$) but narrowly fails convergence (-0.04). The admissible set is empty—a rejection of the pair $(\mathscr{M}, \mathscr{R})$, not of C. The honest response is to revise the menu or the network, never the tolerance δ.

DGP D: holdout discipline. Use a lenient selection screen, $\text{Corr}(m, V_c) \geq 0.40$ and $|\text{Corr}(m, V_d)| \leq 0.60$, and a stricter holdout screen, $|\text{Corr}(m, V_d)| \leq 0.20$ and a group gap of at least 0.25. Selection admits m_{true}, m_{weak}, and m_{contam}; on holdout, m_{contam} fails the tight discriminant test (-0.35) while the other two pass, giving a construct-guided pass rate of $2/3$. An anchor-fit baseline that keeps the two measures with the largest convergent correlations retains m_{true} and m_{contam}, of which only one survives holdout: pass rate $1/2$. In this population-limit example, construct-guided selection beats single-anchor selection on held-out validity—the direction of the falsifiable claim.

DGP E: theory misspecification (a divisible construct). Suppose the construct is initially treated as unitary but the menu actually contains two coherent families with opposite nomological relations: four measures load 0.80 or 0.60 on a subconstruct C_1 and four load 0.80 or 0.60 on C_2 (with $C_1 \perp C_2$), plus noise. Let the convergent criterion be $V_c = 0.80C_1 - 0.20C_2 + \varepsilon$ and the outcome $Y = 0.60C_1 - 0.50C_2 + \varepsilon$, both standardized. A screen $\text{Corr}(m, V_c) \geq 0.39$ admits the C_1 family (population convergent correlations 0.64 and 0.48) and rejects the C_2 family (-0.16 and -0.12). The unrestricted image over $\mathscr{M}$ is $[-0.40, 0.48]$—broad and structured, with a coherent negative block from the rejected family; the conditional image is $\mathscr{B}^* = [0.36, 0.48]$, narrow and positive. Both objects are correct under their own questions: the conditional set is valid inference under $\mathscr{R}$, and the structured disagreement in the full menu is exactly the evidence that C may be divisible. The framework decides which measures may speak under the declared argument; it does not decide whether the rejected family is invalid measurement or evidence that C should be split—that decision belongs to the discovery channel (Appendix D), which revises, refreezes, and tests on new data.

Empirical companion: validity-argument sensitivity. The network grid of Appendix G instantiates both regimes on real data: the trust application behaves like this misspecification case—the qualitative conclusion depends on where the education floor sits—while the patience application behaves like the clean case.

DGP F: reliability attenuation (measurement precision). Hold latent validity constant and vary only the psychometric precision of two otherwise equivalent composites. Let C be standardized with a criterion B at $\text{Corr}(C, B) = 0.70$, and let A_1 and A_2 be means of five items each loading 0.90 and 0.50 on C (independent errors), giving composite reliabilities $\omega_1 = 0.955$ and $\omega_2 = 0.625$. The observed criterion correlations are $\text{Corr}(A_1, B) = 0.70\sqrt{0.955} = 0.684$ and $\text{Corr}(A_2, B) = 0.70\sqrt{0.625} = 0.553$: identical latent validity, different admissibility. A convergent gate $\text{Corr}(m, B) \geq 0.60$ admits A_1 and rejects A_2 purely because of reliability, and the same attenuation makes A_2 look cleaner on an upper-bound discriminant gate against a slightly contaminated foil ($\text{Corr}(A_2, V_d) = 0.079$ versus 0.098 for A_1 when $\text{Corr}(V_d, C) = 0.10$). The lesson is that admissibility can depend on the machinery of score construction, not only on substantive construct differences—the motivation for the measurement-precision layer on the roadmap (Section 6.7), and a caution the paper's own reliability-naive composite $z(\text{Q13}) + z(\text{Q14})$, dragged under the bar by Q14, illustrates empirically.

A finite-sample engine check under the shipped implementation reproduces the qualitative structure of DGPs A–F; run identifiers are recorded in the companion repository.

C Benchmarks as measurement menus

A benchmark maps model behavior into a score interpreted as a capability. If several task families purport to measure the same capability, they form a measurement menu, and construct-identified benchmark inference would define the capability and intended use, declare the benchmark families, specify convergent and discriminant evidence, reserve held-out task families, and report model comparisons over the admissible benchmark set [55, 4]. This perspective does not solve contamination or benchmark governance [54], but it clarifies the scientific object: a leaderboard score is an operationalization, not the capability itself. The motivation is empirical: the AI Index reports capability-benchmark reporting by leading developers as near-universal while responsible-AI benchmark reporting remains uneven [51].

D Validation gates in automated social science

Language models can help generate hypotheses, propose coding schemes, construct candidate measures, and simulate respondents [40]. The design question is which commitments must remain inspectable when the workflow becomes agentic. Automation may propose hypotheses, generate candidates, compute frozen scores, and execute declared restrictions; it must not silently choose the construct definition, rewrite thresholds after observing outcomes, or validate one model with another model's correlated outputs as though they were independent evidence. Interventions in LLM-simulated-user experiments can induce user drift, motivating negative-control outcomes as a diagnostic [39]; prompt-engineering and fine-tuning repairs must be distinguished from statistical calibration using auxiliary human data under explicit assumptions [30]; and early case studies find that frontier agents complete engineering work without substantially answering the research questions [35]. The most direct Goodharting failure is training a generator against the known restriction network; the defenses are procedural—freeze $\mathscr{R}$ and τ before final scoring, preserve negative controls, reserve restrictions or units, and keep generator provenance separate from validity evidence. The medium-run opportunity is a community registry of construct definitions, frozen score matrices, researcher-authored networks, and reproducible audits—a way to make disagreements legible rather than a closed "verifiable reward" for autonomous science.

E Open-weight LLM measurement protocol

LLM outputs enter the framework only after they are reduced to declared score columns. For a binary or ordered response prompt, a log-probability scalarization uses the model-implied probability assigned to prespecified answer tokens. The prompt template, model identity, quantization, tokenizer, decoding settings, answer-token mapping, and scalar reduction are part of the measurement function m and must be frozen with the score provenance. Temperature-zero decoding improves reproducibility but does not establish validity; open weights improve auditability but do not confer scientific standing.

The scores in Section 5 and Appendix F use Meta-Llama-3.1-8B-Instruct Q8_0 and Phi-4-mini Q8_0, GGUF-pinned by checksum, scored with llama.cpp at temperature 0 and seed 20260815. Both models see identical prompt bytes: a silicon-sampling wrapper (typical adult, or woman/man aged in a named band, living in the English country name) plus the published GPS qualitative item and, for trust, the WVS Q57 stem. The country score is the mean of a 0–10 expected rating and a binary $P(B)$, each mapped to the unit interval. Table 7 states the required provenance record.

Table 7: Minimum adapter card for an LLM measurement function.

Field	Required record
Construct prompt	Exact text and response format.
Model	Repository/model identifier and pinned artifact hash.
Runtime	Inference engine, quantization, and tokenizer.
Decoding	Temperature, seed, and any sampling parameters.
Scalarization	Exact map from token probabilities/outputs to the score column.
Provenance	Whether prompts/models were screened before the menu freeze; all exclusions.
Independence	Any overlap between training data, validity auxiliaries, and downstream outcomes.

F Country profiles (underpowered complete application)

These profiles use the same frozen networks and the same two cheap columns as Section 5. They are not the powered application: $n = 41$ and $n = 35$ do not support a units-split headline.

Both open-weight patience columns survive the education and risk-discriminant screen with leftover monotonicity, joining GPS patience. Llama's education slack is $+0.027$ (boundary); Phi's is $+0.071$. WVS Q13, Q14, and the composite still fail education. The plug-in range is $[0.402, 0.565]$ (GPS 0.402, Llama 0.447, Phi 0.565), with coverage $[0.157, 0.674]$ and empty-replicate rate 0.089. Without the cheap columns the set was the GPS singleton 0.402 with empty-replicate rate 0.297.

On trust, Llama survives rule of law and the gini sign restriction (slacks $+0.475$ and $+0.197$) and leftover education ($+0.221$), with $\beta = 0.481$. Phi fails gini (slack -0.158) despite passing rule of law. GPS trust still fails rule of law. The admissible set is $\{\text{Q57}, \text{in}, \text{out}, \text{composite}, \text{Llama}\}$ with range $[0.107, 0.481]$ and coverage $[-0.016, 0.699]$ that still crosses zero. The two models are not two methods: they disagree on the inequality bar, and only one of them may speak.

Table 8: Country appendix (`cvprofiles` 3.0.2, seed 20260814, $n_{\text{boot}} = 1000$). Country β is standardized OLS on log GDP per capita with an education control.

Profile	n	$\mathscr{M}^*$	$[L, U]$	**Coverage**	**Empty-rep.**
Patience, country	41	{GPS, Llama, Phi}	$[0.402, 0.565]$	$[0.157, 0.674]$	0.089
Trust, country	35	{Q57, in, out, composite, Llama}	$[0.107, 0.481]$	$[-0.016, 0.699]$	0.001

G Validity-argument sensitivity grid

This appendix executes the layer named in Section 3.5: seven substantively defensible validity arguments run on the same frozen, country-demeaned cell scores and estimands as the powered application (package 3.0.2 primitives, $\delta = 0$, no bootstrap—this layer is not sampling uncertainty). All variants hold the leftover holdout monotonicity restriction ($\theta = 0.15$) fixed and vary only the select-stage education floor; the last row is the named unrestricted special case ($\mathscr{R} = \emptyset$). The frozen confirmatory network is row $\mathscr{R}_0$ and reproduces Table 3 exactly.

Table 9: Validity-argument sensitivity on the powered cell surface. θ_{edu} is the select-stage education floor ($\text{Corr}(m, \text{q275_cell}) \geq \theta_{\text{edu}}$); all other structure held at the frozen values. Ranges are the plug-in image $[L, U]$ over $\mathscr{M}^*(\mathscr{R}_i)$, rounded to three decimals. Unrestricted patience $[-0.326, 0.556]$ is Q14 / Phi on {Q13, Q14, composite, Llama, Phi} (noise interior). Unrestricted trust $[-0.317, 0.318]$ is Phi / composite on {Q57, in-group, out-group, institution, composite, Llama, Phi} (noise interior).

Network	θ_{edu}	$\mathscr{M}^*(\mathscr{R}_i)$	$[L, U]$
Patience, cells ($n = 480$)			
$\mathscr{R}_0$ (frozen)	0.20	Q13, Llama, Phi	$[0.245, 0.556]$
$\mathscr{R}_1$	0.10	Q13, Llama, Phi	$[0.245, 0.556]$
$\mathscr{R}_2$	0.15	Q13, Llama, Phi	$[0.245, 0.556]$
$\mathscr{R}_3$	0.25	Q13, Llama, Phi	$[0.245, 0.556]$
$\mathscr{R}_4$	0.30	Llama, Phi	$[0.417, 0.556]$
$\mathscr{R}_5$	0.40	Llama, Phi	$[0.417, 0.556]$
$\mathscr{R}_6$ (unrestricted)	—	Q13, Q14, composite, Llama, Phi	$[-0.326, 0.556]$
Trust, cells ($n = 480$)			
$\mathscr{R}_0$ (frozen)	0.20	Phi	$[-0.317, -0.317]$
$\mathscr{R}_1$	0.10	Llama, Phi	$[-0.317, 0.179]$
$\mathscr{R}_2$–$\mathscr{R}_4$	0.15–0.30	Phi	$[-0.317, -0.317]$
$\mathscr{R}_5$	0.40	$\emptyset$	—
$\mathscr{R}_6$ (unrestricted)	—	Q57, in, out, inst., composite, Llama, Phi	$[-0.317, 0.318]$

Two readings travel with the table. First, the patience homogeneity is not an artifact of one boundary choice: the screened range is identical for every floor a reasonable author might defend up to 0.25, and remains strictly positive under floors that admit only the open-weight columns. Second, the trust result is genuinely validity-argument-dependent: at a lenient floor a second cheap score with positive association to GPS trust enters, moving the range across zero, while a very strict floor produces the empty set. The screen-versus-estimand conflict reported in Section 5.4 is therefore itself conditional on where the education bar sits—which is precisely the information the contrast between the unrestricted and conditional images is designed to expose, and why the grid accompanies every future freeze rather than replacing the τ-grid.

H Profiling the inadmissible set

A validity argument is a selection rule, and a selection rule is tested by what it selects *out*. This appendix profiles $\mathscr{M} \setminus \mathscr{M}^*$ across all four frozen runs: which restrictions reject, whether rejections cluster, and whether the rejected columns carry criterion-relevant signal. Four findings survive inspection; all numbers come from the same freezes as Tables 3–8 (package 3.0.2 primitives, no re-scoring).

Finding 1: rejection is restriction-concentrated, not scattered. At the powered cell resolution every human rejection comes from the single education floor—rejection rates are 1.0 for conv_edu and 0.0 for all other select restrictions on the patience menu, and 1.0 versus 0.5 (rule of law versus gini) on the trust country menu. No measure fails by scattering small violations across unrelated implications; each reject has one binding reason. Concentrated rejection supports reading the screen as testing one substantive expectation per measure family, rather than accumulating many weak cuts whose intersection would be uninterpretable.

Finding 2: patience rejects are coherent and estimand-aligned; the screen and the estimand agree. The three rejected patience operationalizations form one tight family (mean inter-reject Spearman $+0.72$ at country level) with uniformly negative development associations, and among human cell measures the rank correlation between the education slack and β is exactly $+1.0$: failing the gate and pointing against the estimand carry the same information. For patience, screening removes nothing the estimand wanted; the A1/A2 concern (Appendix B, DGP E) does not bind on this application.

Finding 3: trust rejects are coherent and estimand-*anti*-aligned; the screen and the estimand disagree systematically. Among human trust-cell measures the correlation between the education slack and β is -0.80 (Spearman/Pearson): the measures that co-move most strongly with GPS trust inside countries ($\beta = 0.21$–0.32) fail the education floor hardest, while the unique passer points negative. This is not a cheap-column artifact—the statistic is computed on human measures only—and it upgrades the Section 5.4 disagreement from a singleton conflict to a structural property of the menu: within countries, the declared education implication and criterion recovery are opposed, which is precisely the pattern in which a failed restriction should be read as evidence about the network rather than about the measures.

Finding 4: the validated instrument is the outlier in its own construct family. At country level, GPS trust correlates with rule of law at $r = +0.12$ while Q57, in-group, and out-group sit between 0.48 and 0.71 (institution $r = -0.12$; composite $r = 0.40$), and GPS trust also fails the leftover education check. The experimentally validated behavioral item behaves unlike {Q57, in-group, out-group}—so the country-level verdict that "GPS trust fails rule of law" is better read symmetrically: either the behavioral measure records something those facets do not, or the institutional expectations in $\mathscr{R}$ apply to belief-trust rather than behavioral trust. The framework cannot decide between those readings; profiling the inadmissible set is what makes the choice visible instead of buried under a per-instrument pass/fail.

Together the findings sharpen what the framework promises. Screening is not lossless compression of the menu: for patience it is nearly free (aligned screen and estimand), while for trust it buys admissibility at the price of criterion relevance—and the price is now measured, not merely conceded. The discovery channel of Appendix D takes Finding 3 as its concrete input: a revised network for within-country trust (for example, replacing the pooled education implication with a level-appropriate one) is exactly the kind of hypothesis that should be frozen and tested on fresh data.

I Leave-one-measure-out

Menu-curation check on frozen slacks and β (cell profiles). Dropping both cheap patience columns leaves the Q13 singleton (already in the main text). Dropping Phi from the trust menu empties $\mathscr{M}^*$.

Table 10: Leave-one-measure-out on frozen cell slacks. Unrestricted ranges are the image of β over the remaining menu.

Drop	$\mathscr{M}^*$	$[L,U]$	Unrestricted
Patience cells			
Q13	Llama, Phi	$[0.417, 0.556]$	$[-0.326, 0.556]$
Q14	Q13, Llama, Phi	$[0.245, 0.556]$	$[-0.027, 0.556]$
composite	Q13, Llama, Phi	$[0.245, 0.556]$	$[-0.326, 0.556]$
Llama	Q13, Phi	$[0.245, 0.556]$	$[-0.326, 0.556]$
Phi	Q13, Llama	$[0.245, 0.417]$	$[-0.326, 0.417]$
noise	Q13, Llama, Phi	$[0.245, 0.556]$	$[-0.326, 0.556]$
Trust cells			
Q57	Phi	$[-0.317, -0.317]$	$[-0.317, 0.318]$
in-group	Phi	$[-0.317, -0.317]$	$[-0.317, 0.318]$
out-group	Phi	$[-0.317, -0.317]$	$[-0.317, 0.318]$
institution	Phi	$[-0.317, -0.317]$	$[-0.317, 0.318]$
composite	Phi	$[-0.317, -0.317]$	$[-0.317, 0.285]$
Llama	Phi	$[-0.317, -0.317]$	$[-0.317, 0.318]$
Phi	$\emptyset$	empty	$[0.015, 0.318]$
noise	Phi	$[-0.317, -0.317]$	$[-0.317, 0.318]$

J Level-dependence diagnostic for the cross-resolution reversal

Section 5.4 noted that measurement noninvariance could compete with the substantive reading of the pooled-versus-demeaned reversal. This appendix runs the strongest test available on derived frames—no item-level microdata, so classical multi-group invariance testing is out of scope—and asks whether the *measurement function* or the *expectation* carries the level-dependence (480 cells, 42 countries; package 3.0.2 inputs).

Design. Three comparisons per measure: variance decomposition of each cell column into between- and within-country shares; the association between measure and education auxiliary at country level (country means), pooled cells, and within-country (demeaned) cells; and country-slope heterogeneity of the measure–education association inside countries.

Result 1: thrift's reversal is expectation-level-dependence, not item DIF. Q13's education association flips sign across levels (-0.33 country, -0.17 pooled, $+0.25$ within), but its within-country slope is positive in 79% of the 42 countries (mean $+0.21$, median $+0.32$)—the within-country expectation holds nearly everywhere rather than being manufactured by aggregation. Q13–Q14 Pearson co-movement, named by frame: 0.52 on the country GPS/WVS join ($n = 41$); 0.48 on cell country-means ($n = 42$); 0.35 on pooled cells ($n = 480$); -0.23 on demeaned cells. The reversal therefore lives in how the education implication binds at different levels, not in a single cross-level r; classical item-level DIF remains the stronger unrun test.

Result 2: perseverance is uniformly negative—a stable property, not an artifact. Q14's education association is negative at every level (-0.59 country, -0.53 pooled, -0.36 within) and its within-country slope is negative in 88% of countries (mean -0.38). Its anti-alignment is a level-robust feature of the measure, consistent with treating perseverance as substantively distinct from time preference rather than as a resolution-dependent artifact.

Result 3: trust facets show no structural reorganization either. Every facet's association with GPS trust attenuates smoothly from country to within-cell level (0.27–0.47 down to 0.05–0.32, ordered identically at both levels), and the composite's education association declines monotonically ($+0.15 \to +0.08 \to -0.27$) without sign reorganization among facets.

Verdict for this application: no evidence of structural noninvariance in the derived frames—the cross-resolution findings survive their most plausible psychometric rival on the available evidence, with item-level invariance testing flagged as the remaining stronger test (Section 6.7).

K Response to the psychometrician's review

An earlier draft received a detailed review from a psychometrician; this revision is built around it, and we thank the reviewer publicly here. The comments clustered into four threads, each of which changed the paper: (A) the robustness claim was conditional on the validity argument and read as one-way adjudication; (B) measurement precision and reliability were absent from both the screen's inputs and its interpretation; (C) cross-resolution reversals could reflect measurement noninvariance rather than construct relations; (D) the paper claimed too much of the validity ecosystem for what it operationalizes. This appendix records where each thread landed, so that readers—and the reviewer—can audit the response without diffing drafts.

Thread A: conditionality and the reciprocal channel (comments on the core claim, M*, thresholds). Adopted throughout. Robustness language was replaced by heterogeneity-of-results framing; survivors are written $\mathscr{M}^*(\mathscr{R})$ to make their dependence on the declared argument visible; every headline is conditional on restriction validity, now a formal main-text assumption (Assumption 3.1); unrestricted and screened images are always reported side by side. The confirmation-bias concern behind the A1/A2 example was taken seriously enough to be tested twice: DGP E (Appendix B) simulates it, and Appendix G executes the competing-networks layer the comments proposed—seven defensible arguments on one frozen surface, showing patience homogeneity survives argument variation while trust does not. Appendix H profiles $\mathscr{M} \setminus \mathscr{M}^*$ directly and finds the reviewer's concern empirically confirmed on one construct: among human trust-cell measures the correlation between education slack and estimand alignment is -0.80, so screening there buys admissibility at a measured price in criterion relevance—the finding is reported as such and handed to the discovery channel, not tuned away. The "validity decides who may speak" metaphor was withdrawn; rejected measures are treated as evidence throughout.

Thread B: precision and reliability. Adopted as a named roadmap commitment with its motivating analysis already in the paper: reliability attenuation has its own DGP (Appendix B, DGP F), Section 6.7 specifies the planned extension (optional per-measure reliability metadata; disattenuated slack and β variants reported alongside naive ones), and auxiliaries are explicitly flagged as fallible when they are themselves construct scores. Not yet implemented in the engine; deliberately not promised as done.

Thread C: noninvariance. Adopted and executed to the limit of available data. Appendix J runs the strongest test the derived frames permit—three-level association comparisons plus within-country slope heterogeneity—finds no structural reorganization of any measure across levels, and locates thrift's reversal in level-dependent binding of the education implication rather than item functioning. Item-level multi-group invariance testing remains flagged as the stronger unrun test wherever microdata access permits.

Thread D: positioning within the validity ecosystem. Adopted. The paper now claims criterion-related validity assessment across competing measures—relations-to-other-variables evidence under a declared argument—rather than construct validation generally; keywords, abstract, and Section 2 carry the narrowed claim; the Campbell–Fiske special case is explicitly correlational-criteria nesting, not latent-variable MTMM recovery.

Where we differ, and why. Two suggestions were partially declined, with reasons. First, renaming the pipeline around "heterogeneity" alone: we kept *construct-identified inference* as the frame because the partial-identification register is what makes two-set reporting constitutive rather than decorative, but every interpretive sentence now carries the heterogeneity reading the reviewer preferred. Second, fully reciprocal theory revision inside the confirmatory lane: the paper keeps revision in a frozen-and-retested discovery channel (Appendices D, H) because folding it into the screen would recreate the circularity the same review warned against; we note the reviewer's own framing anticipates this in requiring revised contracts to face new data. We record these declines so the disagreement can be adjudicated by readers rather than buried.

Remaining gaps, stated plainly: the precision layer is unimplemented; the competing-networks grid covers threshold variation within one network family, not between-family alternatives; item-level invariance testing awaits licensed microdata; and the within-country trust revision motivated by Finding 3 of Appendix H is specified but not yet frozen or fielded.

L Synthesis of contributions and empirical grounding

Table 11 summarizes the scope and depth of our methodological and substantive contributions, framing construct-identified inference as a formal bridge across psychometrics, multiverse reporting, and partial identification. Figure 3 details the synthesis of these three traditions in response to the AI measurement shock.

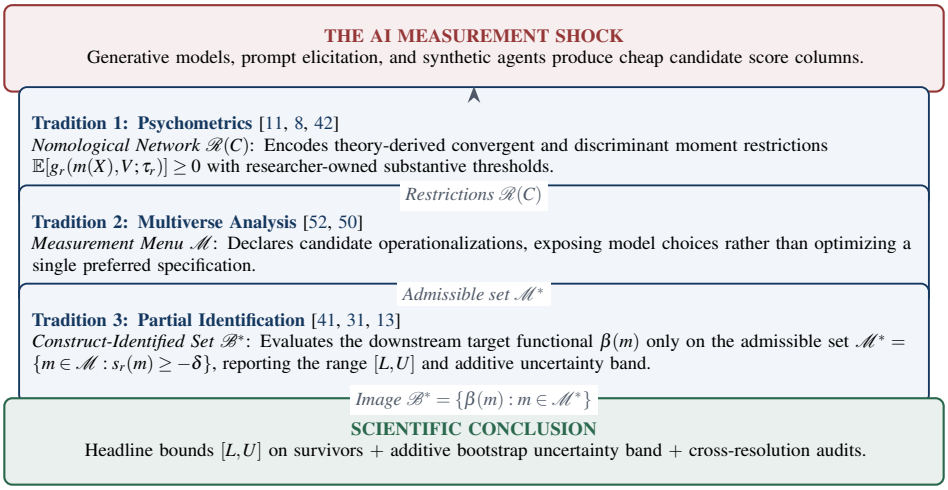


Figure 3: The three traditions composed by construct-identified inference. Nomological restrictions from psychometrics screen candidate operationalizations from a declared multiverse, and partial identification propagates the remaining measurement choice into downstream scientific conclusions.

Table 11: Summary of methodological contributions and their empirical grounding in this paper.

Claimed Contribution	Methodological Anchor	Key Literature	Empirical Grounding in This Paper
I. Unrestricted Multiverse Recovery	Setting $\mathscr{R} = \emptyset$ recovers the raw specification curve as a named special case of $[L, U]$.	Simonsohn et al. [50]; Steegen et al. [52]	On the present 7-measure country patience menu (GPS, Q13, Q14, composite, Llama, Phi, noise), $\mathscr{R} = \emptyset$ returns $[-0.219, 0.565]$ (endpoints composite / Phi), so raw multiverse reporting still sign-switches on development.
II. Formal MTMM Matrix Recovery	Campbell–Fiske convergent (τ_{conv}) and discriminant (τ_{disc}) cells are ordinary corr_min / corr_zero restrictions.	Campbell & Fiske [8]; Falk et al. [17]	The 2×2 matrix of patience and trust empties: WVS Q57 tracks GPS patience ($r = 0.825$; partial $r = 0.706$) far more than GPS trust ($r = 0.284$).
III. Sampling Fragility & Uncertainty Bands	Selection on sample moments is data-dependent; units bootstrap bands and empty-replicate rates are reported additively alongside $[L, U]$.	Imbens & Manski [31]; Stoye [53]	Cheap patience columns join GPS at country level ($[0.402, 0.565]$, empty-rep. 0.089); coverage remains wide.
IV. Resolution Shifts & Ecological Confounds	Testing across macro (country) and within-country demographic cell resolutions reveals whether proxy failure is construct mismatch or aggregation bias.	Falk & Hermle [19]; Messick [42]	WVS Thrift (Q13) fails at country level (education association -0.33) but survives within-country demeaning ($+0.25$, $\beta = 0.245$), while Perseverance (Q14) is education-anti-aligned at every level (-0.59 country, -0.53 pooled, -0.36 within).
V. Construct Selection vs. Estimand Co-movement	Predeclared nomological gates and criterion-recovery estimands can cleanly pull apart without invalidating the framework.	Kane [32]; Borsboom et al. [6]	Human trust cells remain empty under the education gate; only Phi is admitted and its β versus GPS trust is -0.317, while multi-item facets still co-move with GPS trust ($r = 0.21$–0.32).