

The Proxy Presumption: From Semantic Embeddings to Valid Social Measures

Baishi Li¹, Ta Yu¹, Kelvin J.L. Koa^{1,2*}, Ke-Wei Huang^{1,2}

¹National University of Singapore, ²Asian Institute of Digital Finance

✉: baishili@u.nus.edu, ta.yu@u.nus.edu, kelvin.koa@u.nus.edu, dishkw@nus.edu.sg

Abstract

Natural Language Processing is rapidly evolving into a primary instrument for Computational Social Science, with researchers increasingly using embeddings to measure latent constructs such as *novelty*, *creativity*, and *bias*. However, this transition faces a fundamental validity challenge: the “Proxy Presumption,” or the reliance on geometric properties (e.g., cosine distance) as direct measures of social concepts. We argue that without explicit validation, unsupervised representations remain entangled mixtures of the target construct (C) and confounding attributes (Z) like topic, style, and authorship. To bridge the gap between semantic embeddings and valid social measures, we introduce the Construct Validity Protocol (CVP). Drawing on causal representation learning and psychometrics, the CVP offers a rigorous pipeline from conceptualization to quantitative verification. We further propose Counterfactual Neutralization, a novel method using LLMs to reduce confounding in embedding space. By providing a standardized Validity Suite—including tests for discriminant, incremental, and predictive validity—this work offers the community a toolkit to transform heuristic proxies into robust, scientifically defensible instruments.

1 Introduction

The core ambition of contemporary Natural Language Processing (NLP) has expanded beyond predicting the next token; we are increasingly tasked with measuring the social world. Recent literature has pioneered the quantification of abstract social constructs, proposing computational metrics for concepts such as *novelty*, *creativity*, and *bias* (Merrill et al., 2024; Lee et al., 2024; Bang et al., 2024). This transition marks a pivotal moment: NLP is evolving from an engineering discipline into a pri-

mary measurement instrument for Computational Social Science (CSS).

However, this ambition introduces a methodological challenge: the gap between **theoretical constructs** (what we intend to measure) and **computational proxies** (what we implement). In the absence of established ground truth for latent social variables, the field faces a risk we term the *Proxy Presumption*: a convenient geometric property—most commonly cosine similarity in embedding space—is renamed as a theoretical construct without sufficient validation (Caliskan et al., 2017; May et al., 2019; Bolukbasi et al., 2016; Garg et al., 2018; Kozłowski et al., 2019). We emphasize that this is a recurring pattern across a substantial and influential subset of NLP/ML measurement work rather than an isolated mistake, while recognizing meaningful variation in how rigorously different papers validate their proxies. While vector distance captures semantic divergence, equating it directly with “innovation” or “creativity” assumes a strong isomorphism between embedding space geometry and social reality that may not hold.

While measurement validity is a general issue in machine learning, the proxy presumption is especially acute in NLP because text is the primary unstructured medium for computational social science, and major nuisance dimensions (Z) in text are inherently linguistic (register, dialect, pragmatics, style). Moreover, many interventions that can reduce confounding are text-native, including counterfactual rewriting and language-aware debiasing methods (Elazar and Goldberg, 2018; Ravfogel et al., 2020; Blodgett et al., 2020; Sap et al., 2019).

This paper argues that such “measurement by renaming” can be formalized as a non-identification problem: unsupervised embeddings compress multiple generative factors into a single representation, entangling the target construct (C) with nuisance attributes (Z) such as topic, author style, register, or length (Locatello et al., 2019; Schölkopf et al.,

*Corresponding author.

2021). Without explicit disentanglement or design-based controls, standard metrics may primarily capture method variance rather than the intended social phenomenon (Zhou et al., 2022). For example, a high cosine distance may reflect a genuine conceptual break, or it may simply reflect a vocabulary shift. Even with perfect disentanglement, the current practice still suffers from the mis-specification issue: the function (such as cosine similarity) chosen to represent the new concept may not be the most suitable function that represents the functional relationship between C and the document, excluding the influences from confounding covariates Z .

To address this challenge and support the maturation of CSS, we introduce the **Construct Validity Protocol (CVP)**. Adapted from standards in psychometrics and causal representation learning (Bengio et al., 2013; Locatello et al., 2019; Schölkopf et al., 2021), the CVP provides a roadmap for validating social variables in NLP. Our contributions are fourfold:

1. **A non-identification result:** We formalize why unsupervised metrics cannot reliably recover latent constructs without explicit assumptions or interventions (Locatello et al., 2019; Schölkopf et al., 2021).
2. **Methodological solutions:** We propose **Counterfactual Neutralization** and connect it to existing debiasing/disentanglement tools such as adversarial removal and nullspace projection (Elazar and Goldberg, 2018; Ravfogel et al., 2020).
3. **A Validity Suite:** We define a standardized suite of evaluations (stability/reliability, convergent validity, discriminant and incremental validity, known-groups checks, and criterion-related evidence) to test whether a proxy tracks the intended construct beyond nuisance dimensions.
4. **A forensic literature review:** We analyze recent publications (2021–2025) to document the reuse of nearly identical similarity-based instruments for distinct, sometimes contradictory, constructs.

Scope and positioning. This paper is a *position and synthesis* contribution: we propose CVP as a community-facing reporting protocol for measurement identification, not a final or exclusive standard. We present CVP as a starting point that

should be iterated as evidence accumulates and as NLP-specific failure modes are better understood, analogous in spirit to evolving reporting guidelines in adjacent areas (e.g., Data Statements for NLP; Bender and Friedman (2018)).

Our goal is not to discourage the measurement of complex social phenomena, but to provide tools that render such measurements scientifically defensible. In Computational Social Science (CSS), these constructed variables are often critical inputs for downstream causal inference or predictive tasks. However, if the variable construction itself lacks validity, no causal ML methods can rigorously establish downstream causality.

2 Related Literature: The State of Social Measurement

Our work connects three research traditions that are often cited separately but rarely integrated in NLP: (i) measurement theory and construct validity from the social sciences, (ii) causal representation learning and identifiability, and (iii) recent NLP methodology debates about what models and annotations actually measure.

2.1 Measurement Theory and Construct Validity.

Measurement theory distinguishes *constructs* (theoretical concepts) from *measures* (observable indicators). As codified in standard social science guidelines (DeVellis, 2016; Adcock and Collier, 2001), validation is not a definitional act but an empirical one: researchers must demonstrate that a measure behaves as predicted within a “nomological network,” correlating with related concepts (convergent validity) while remaining distinct from nuisances (discriminant validity). This rigorous framework is increasingly relevant to NLP as the field targets complex social variables. A growing body of work now operationalizes constructs through explicit *domains of observables*—for example, decomposing persuasion into specific resistance strategies or mapping social norms to defined interactional patterns (Vijjini et al., 2024; Huang and Yang, 2023). These studies exemplify the “construct-to-measure” pipeline we formalize, standing in contrast to scalar metrics that prioritize prediction without separating signal from confounding noise.

2.2 Validity in NLP: Annotation, Bias, and the Limits of Proxying.

A parallel line of work argues that many NLP variables inherit ambiguity from their measurement process. Human label variation is increasingly recognized not as annotator error, but as a meaningful signal of linguistic ambiguity and subjectivity (Pavlick and Kwiatkowski, 2019; Plank, 2022; Davani et al., 2022; Aroyo and Welty, 2015). Furthermore, dataset curation and annotation protocols can embed structural artifacts and degenerate shortcuts into text, yielding variables that appear to measure a target construct while in fact tracking annotation or sampling regularities (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019; Bender and Friedman, 2018; Geiger et al., 2020). Methodological critiques in fairness and social NLP emphasize that quantities like “toxicity” and “bias” are often operationalized as convenient proxies that suffer from construct slippage and normative mismatch, including dialect and genre confounds (Blodgett et al., 2020; Sap et al., 2019). This perspective complements our claim that measurement failures are frequently *method-variance failures*: what appears to be a social construct may largely reflect dataset composition, genre, dialect, or frequency artifacts. Our CVP makes these critiques actionable by requiring explicit confound tests, invariance checks aligned with established NLP evaluation practice (Ribeiro et al., 2020; Gardner et al., 2020), and transparent reporting of what the proxy is sensitive to.

2.3 Causal Representation Learning and Non-Identification.

We ground our framework in Causal Representation Learning (Schölkopf et al., 2021), which establishes a fundamental non-identification result: without explicit structural assumptions or interventions, it is theoretically impossible to disentangle latent causal factors from observational data alone. We apply this constraint to social measurement: extracting a valid construct (C) from text requires actively modeling and neutralizing nuisance factors (Z), rather than assuming that an unsupervised embedding will spontaneously isolate them.

Beyond representation learning, a complementary literature studies when text-derived variables can support downstream causal inference (Egami et al., 2022) and how to use imperfect surrogates for downstream inference via design-based estima-

tors (Egami et al., 2023). Learned-proxy methodology in political science further emphasizes that post-hoc correlation is insufficient for testing causal theories with learned proxies (Knox et al., 2022). We position CVP as addressing the measurement-validity prerequisite that these downstream frameworks typically assume.

3 The Impossibility of Unsupervised Inverse Construction

Current computational social science often relies on the assumption that a specific latent social construct C (e.g., a hate sentiment score) can be recovered directly from observed documents D using a fixed proxy function $f(D)$. We argue that this formulation constitutes an *ill-posed inverse problem* formally analogous to the identification problem in causal inference.

3.1 Case I: Single-Document Measurement

Let the generation of a document $D \in \mathcal{D}$ be modeled as a stochastic process dependent on two disjoint sets of latent factors: a **scalar** target construct $c \in \mathbb{R}$ and a vector of nuisance attributes $\mathbf{z} \in \mathbb{R}^k$ (e.g., *Topic, Length, Authorship*). We model the document generation process as a draw from a conditional probability distribution parameterized by θ .

$$p_{\theta}(D | c, \mathbf{z}) \quad (1)$$

The standard measurement pipeline attempts to invert this process:

1. **Representation Learning:** An encoder $E : \mathcal{D} \rightarrow \mathbb{R}^d$ maps the text to an embedding space \mathbf{e} .
2. **Proxy Construction:** A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is applied to estimate the scalar construct: $\hat{c} = f(\mathbf{e})$.

Ideally, $f(E(D))$ should recover c . However, as established in the representation learning literature (Locatello et al., 2019; Schölkopf et al., 2021), this recovery is theoretically impossible without structural assumptions.

Proposition 1 (Rotational Ambiguity). *Let the joint latent space be $\mathbf{h} = [c; \mathbf{z}]$ with a factorized isotropic Gaussian prior $p(\mathbf{h}) = \mathcal{N}(0, I)$. For any unsupervised objective that maximizes the likelihood of the data $p(D)$, the learned representation is identifiable only up to an arbitrary orthogonal rotation. Consequently, the scalar dimension c is inextricably mixed with the nuisance vector \mathbf{z} .*

Proof Sketch. Following (Locatello et al., 2019), consider the marginal likelihood $p(D) = \int p_\theta(D | \mathbf{h})p(\mathbf{h})d\mathbf{h}$. Let R be any arbitrary orthogonal matrix ($R^\top R = I$). We define a transformed latent space $\mathbf{h}' = R\mathbf{h}$. Due to the isotropy of the Gaussian prior, the density is rotation-invariant: $p(\mathbf{h}') = p(\mathbf{h})$.

Now, consider an alternative generator defined by $G'(\mathbf{x}) = G(R^\top \mathbf{x})$. If we assume the data generation process uses latent variables \mathbf{h}' and generator G' , the observation is $D = G'(\mathbf{h}') = G(R^\top(R\mathbf{h})) = G(\mathbf{h})$. Thus, the model (G', \mathbf{h}') yields the exact same observational distribution $p(D)$ as the original model (G, \mathbf{h}) .

Since the unsupervised objective depends only on $p(D)$, it cannot distinguish between the entangled basis \mathbf{h}' (where the first dimension is a linear combination of c and \mathbf{z}) and the disentangled basis \mathbf{h} . In other words, any $\mathbf{h}' = \alpha c + \beta^\top \mathbf{z}$ is a possible latent factor vector. Thus, c is not identified.

Implication. This proves that statistical independence does not imply disentanglement (Schölkopf et al., 2021). Even if $c \perp \mathbf{z}$ in the real world, the learned embedding \mathbf{e} can arbitrarily rotate the basis such that the dimension used for measurement is a mixture of signal and noise. Thus, any proxy $\hat{c} = f(\mathbf{e})$ suffers from intrinsic method variance. We will illustrate this claim with an example in the two-document case.

3.2 Case II: Two-Document Measurement

Constructing variables from pairwise document embeddings introduces an additional layer of entanglement. Consider measuring the relationship between two documents, such as *Scientific Novelty* (C), defined as the distance between a new paper D_i and prior work D_j . The key nuisance is topical overlap: both documents may share a dominant topic (e.g., “Deep Learning”), which can account for most surface-level lexical and semantic mass. Standard practice then applies a similarity or distance function (e.g., cosine similarity) to the *full* embeddings, implicitly allowing this nuisance component to dominate the measurement.

Even if we assume an oracle encoder E that produces a perfectly disentangled concatenation $\mathbf{e} = [c \oplus \mathbf{z}]$, where c encodes the concept of interest and \mathbf{z} encodes nuisance variation, similarity-based measurement can still fail for two reasons:

(I) **Target mismatch.** The metric of interest should be applied to the concept embed-

dings (c_i, c_j) , but is typically applied to the full vectors $(\mathbf{e}_i, \mathbf{e}_j)$, allowing variance in \mathbf{z} to dominate.

(II) **Metric indeterminacy.** Even if \mathbf{c} were perfectly isolated, there is no theoretical guarantee that a particular geometric function (e.g., cosine similarity versus Euclidean distance) is the correct proxy for an abstract construct. Determining the appropriate functional form requires empirical validation (Section 4).

We now illustrate issue (I) by expanding two common choices of metrics applied to the full embedding \mathbf{e} .

Cosine similarity (normalization effect).

$$\cos(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{c}_i \cdot \mathbf{c}_j + \mathbf{z}_i \cdot \mathbf{z}_j}{\sqrt{\|\mathbf{c}_i\|^2 + \|\mathbf{z}_i\|^2} \sqrt{\|\mathbf{c}_j\|^2 + \|\mathbf{z}_j\|^2}} \quad (2)$$

When $\|\mathbf{z}\| \gg \|\mathbf{c}\|$, the denominator is dominated by the nuisance component, so high cosine similarity largely reflects “same topic” rather than the intended construct (e.g., “novel conceptual contribution”).

Euclidean distance (additive decomposition issue).

$$\|\mathbf{e}_i - \mathbf{e}_j\|^2 = \|\mathbf{c}_i - \mathbf{c}_j\|^2 + \|\mathbf{z}_i - \mathbf{z}_j\|^2 \quad (3)$$

Because the contributions of Δc and Δz are purely additive, a large distance can arise from genuine conceptual deviation (ΔC) or from nuisance shifts such as subfield jargon or topical drift (ΔZ), and the metric alone does not identify which source drives the score.

3.3 Strategies for Alleviation

Because perfect disentanglement from observational text is not guaranteed without additional structure or assumptions, we outline a practical, multi-level strategy that intervenes at three points in the pipeline: (i) the input text, (ii) the learned representation, and (iii) the scoring function. The goal is to approximate measurement validity by reducing dependence on nuisance factors Z while preserving information relevant to the construct C .

Level 1: Input disentanglement (preprocessing D). The most direct intervention is to reduce the dependence of the observed document D on nuisance factors before representation learning. Concretely, we transform $D \rightarrow D'$ to preserve

construct-relevant content while attenuating nuisance variation. **Targeted information extraction (LLM-based):** rather than embedding the full document, one can use an LLM as a constrained extractor to retain only spans that are logically relevant to the construct definition. For example, when measuring *mission hybridity*, one can extract mission statements and strategic objectives while discarding biographical details, formatting artifacts, or boilerplate that primarily contribute to Z . **Stylistic normalization:** because embeddings are sensitive to length, register, and complexity, inputs can be normalized to a canonical form (e.g., comparable length, standardized structure, or neutral tone) to reduce stylistic nuisance components (e.g., Z_{style}). **Entity masking (anonymization):** when the construct is conceptual (e.g., *policy stance*) but nuisance variation includes entity identity (e.g., organization names, demographic markers), embedding models may rely on entity co-occurrence rather than the intended abstract relation. Replacing named entities with placeholders (e.g., [ORG], [PERSON]) can reduce entity-driven shortcuts and encourage reliance on structural and contextual cues.

Level 2: Representation disentanglement (learning in the e space). If the representation itself is entangled, we can impose objectives that explicitly suppress Z while preserving C . **Supervised/adversarial removal:** when labels for nuisance factors are available (e.g., topic IDs, author attributes, domain indicators), adversarial learning can discourage encodings of Z in e (e.g., Adversarial Removal; Elazar and Goldberg, 2018), and projection-based methods can iteratively identify and remove subspaces predictive of Z (e.g., Iterative Nullspace Projection; Ravfogel et al., 2020). **Contrastive learning:** when one can construct paired examples that preserve C while varying Z , contrastive objectives can encourage embeddings to align with construct-invariant features rather than nuisance variation.

Level 3: Methodological disentanglement (designing the scoring function f). When residual entanglement remains, the measurement function itself can be designed to cancel nuisance contributions. **Counterfactual neutralization:** instead of reporting a static score $f(e_{\text{obs}})$, define measurement as a differential relative to a baseline embedding:

$$\hat{C} = f(e_{\text{obs}}) - f(e_{\text{base}}), \quad (4)$$

where e_{base} is the embedding of a counterfactual “neutralized” version of the same text in which the construct signal is minimized (e.g., rewriting to remove stance/novelty claims while preserving topical content). The subtraction aims to remove the baseline contribution attributable to nuisance factors Z , leaving a score more sensitive to the construct-relevant component.

4 The Construct Validity Protocol (CVP)

Construct validity is not merely a quality check; it is a prerequisite for identification. In both social science and machine learning, isolating a target factor from nuisance variation requires additional structure (e.g., design constraints or validation evidence), much like causal representation learning requires structural assumptions to separate latent causes from noise (Locatello et al., 2019; Schölkopf et al., 2021). For embedding-based measurement, the core requirement is that variation in the proxy \hat{C} is attributable to the intended construct C , rather than superficial regularities such as topic, style, or length (Z) (Grimmer and Stewart, 2013; Adcock and Collier, 2001). Without this form of *measurement identification*, downstream inferences are fragile: we cannot interpret effects of a variable that has not been credibly isolated. We therefore propose the *Construct Validity Protocol (CVP)* as a standard operating procedure that moves from an abstract concept to a testable measurement instrument.

4.1 Overview

CVP has three phases. Phase 1 establishes a construct specification and qualitative alignment. Phase 2 emphasizes instrument design choices that reduce entanglement with nuisance factors. Phase 3 reports a structured validity suite—a **Validity Card**—that documents stability and quantitative evidence (convergent, discriminant/incremental, and criterion-related checks).

4.2 Phase 1: Conceptualization

Before any modeling, the researcher should establish **face and content validity** through expert scrutiny (Adcock and Collier, 2001; DeVellis, 2016; Lawshe, 1975; Haynes et al., 1995). This phase clarifies what the construct is *and is not*, and what nuisance dimensions must be neutralized.

Deliverables. Construct map: Provide a concise definition of C and enumerate the primary

nuisance dimensions Z (e.g., domain/topic vocabulary, style/register, length, venue/time). **Facet blueprint:** Specify a short list of facets of C (inclusion/exclusion criteria) and the types of textual evidence expected for each facet. **Exemplar set:** Curate a three-tier anchor set of known high- C , low- C , and borderline/mid-spectrum documents (“known-groups” anchors) that are broadly agreed upon within the research community; borderline anchors are used to probe facet boundaries and decision rules.

Face-validity audit. Run the full measurement pipeline on the exemplar set and inspect the highest- and lowest-scoring cases, then examine borderline anchors to probe decision boundaries. Independent domain experts should judge whether the ranking is plausible and, crucially, which cues appear to drive the scores. Systematic disagreements (including non-ideal false positives and false negatives on the anchors) indicate construct–measure misalignment and should trigger revisions before quantitative validation.

4.3 Phase 2: Operationalization

Rather than defaulting to black-box similarity scores, Phase 2 focuses on designing instruments that reduce entanglement with Z and better isolate C . Section 3 outlines concrete intervention points (input preprocessing, representation learning, and scoring-function design), which can be instantiated as task-specific measurement instruments.

4.4 Phase 3: The Validity Suite

Validity is not a single statistic but an argument supported by multiple, conceptually distinct sources of evidence (Cronbach and Meehl, 1955; Campbell and Fiske, 1959; Messick, 1995; American Educational Research Association et al., 2014). This is especially important for embedding-based measures, where instability and nuisance leakage are common (Grimmer and Stewart, 2013; Gentzkow et al., 2019). We therefore recommend reporting a structured **Validity Card** that documents (at minimum) stability and discriminant/incremental evidence, and (when feasible) convergent and criterion-related evidence.¹

¹Measurement theory distinguishes multiple families of evidence, including content/face evidence, reliability/stability, convergent and discriminant evidence (external structure), and criterion-related evidence (Campbell and Fiske, 1959; Messick, 1995; American Educational Research Association et al., 2014). We prioritize (i) *stability* as a prerequisite and (ii)

Validity Card 1: Reliability / Stability (prerequisite). Even deterministic encoders can yield unstable measures due to preprocessing choices, chunking, or prompt/instrument sensitivity. If small, defensible perturbations materially change the resulting variable, downstream validity claims become specification-dependent (Koo and Li, 2016; Nunnally and Bernstein, 1994).

Recommended perturbations. Swap between close encoder variants; paraphrase prompts; vary neutralization prompts; vary embedding aggregation rules.

What to report. (i) a clearly enumerated perturbation set and the number of variants per document; (ii) *absolute-agreement* reliability via ICC (report the ICC form, e.g., ICC(2,1) or ICC(2, k)) (Shrout and Fleiss, 1979; Koo and Li, 2016). As a rule of thumb, ICC ≥ 0.75 is often interpreted as “good” and ICC ≥ 0.90 as “excellent” (Koo and Li, 2016).²

Validity Card 2: Convergent validity (agreement with an independent measure). Does \hat{C}_i correlate with an independent measure of the same construct (e.g., expert ratings using a rubric that operationalizes C)? This follows multitrait–multimethod logic: measures of the same trait should agree more than measures of different traits (Campbell and Fiske, 1959).

$$C_i^{gold} = \alpha + \beta_{conv} \hat{C}_i + \epsilon_i \quad (5)$$

What to report. (i) reliability of C^{gold} (e.g., ICC or Krippendorff’s α), since convergent evidence is bounded by gold reliability (Spearman, 1904; Koo and Li, 2016); (ii) effect size (standardized β_{conv} or correlation) with confidence interval and a diagnostic plot; and (iii) if possible, out-of-sample convergence via a held-out gold set or cross-validation (Hastie et al., 2009). Effects around 0.10 are often described as small, 0.30 moderate, and 0.50+ large (Cohen, 1988).

Validity Card 3: Discriminant + incremental validity (primary diagnostic). A central threat is that \hat{C} becomes a proxy for topic, style, venue,

discriminant + incremental validity as the most diagnostic for ruling out topic/style surrogacy in embedding pipelines; known-groups and criterion-related checks are recommended complements.

²Pearson correlation can be misleading for stability because it is invariant to shifts and rescalings. ICC for absolute agreement penalizes systematic shifts in level across perturbations.

or time rather than the intended construct (Grimmer and Stewart, 2013; Gentzkow et al., 2019). Discriminant validity asks whether \hat{C} is distinct from nuisance dimensions. Incremental validity asks whether \hat{C} adds explanatory power beyond nuisances for an *external validation label* L_i that directly operationalizes C (often human-coded).

Step 1 (discriminant). Test whether Z mechanically explains \hat{C} :

$$\hat{C}_i = \alpha + \sum_k \gamma_k Z_{k,i} + \epsilon_i. \quad (6)$$

Step 2 (incremental). Test whether \hat{C} adds signal for an external validation label L_i beyond Z :

$$L_i = \alpha + \sum_k \delta_k Z_{k,i} + u_i, \quad (7)$$

$$L_i = \alpha + \beta_{inc} \hat{C}_i + \sum_k \delta_k Z_{k,i} + u_i. \quad (8)$$

What to report. Pre-specify nuisance blocks (topic proxies, length/style measures, venue/field/year fixed effects) and report (i) predictability of \hat{C} from Z (e.g., cross-validated R^2) with block-wise contributions; and (ii) incremental evidence via standardized β_{inc} with confidence interval and out-of-sample ΔR^2 from adding \hat{C} to the nuisance-only model (Hastie et al., 2009; Hunsley and Meyer, 2003). High predictability of \hat{C} from topic/FE blocks is a warning sign of surrogacy; instability of β_{inc} across reasonable nuisance sets is a red flag (Grimmer and Stewart, 2013; Cohen, 1988).

Validity Card 4: Known-groups validity (optimal). Using the Phase 1 exemplar set, test whether the measure separates pre-specified high- C versus low- C groups:

$$\hat{C}_i = \alpha + \tau \mathbb{1}\{i \in \text{High-}C\} + \epsilon_i. \quad (9)$$

What to report. Separation effect (e.g., Cohen’s d with confidence interval) and a distribution plot (e.g., ECDF). As a rule of thumb, $d \approx 0.2$ is small, 0.5 medium, and 0.8 large (Cohen, 1988).

Validity Card 5: Criterion-related / predictive evidence (complement). If C is substantively meaningful, \hat{C} should relate to downstream outcomes in the theorized direction. For validity testing, prioritize interpretability and attribution: a linear specification provides a conservative baseline that tests whether \hat{C} has independent signal beyond strong nuisance controls.

$$Y_{t+k,i} = \alpha + \beta_{pred} \hat{C}_i + \Gamma \text{Controls}_i + \epsilon_i. \quad (10)$$

What to report. Standardized β_{pred} with confidence interval and the incremental explanatory power from adding \hat{C} beyond strong nuisance controls. When possible, include a negative-control outcome Y_{placebo} that should not be predicted by C ; strong prediction of implausible outcomes suggests dataset artifacts (Lipsitch et al., 2010).

4.5 Worked Example

To demonstrate that CVP is operational, we executed it on GoEmotions (Demszky et al., 2020) (43,410 train / 5,426 dev / 5,427 test). GoEmotions is vector-valued; for brevity we instantiate one coordinate (*gratitude*) with $L_i = 1$ iff the human label set contains gratitude. We focus on Cards 1 and 3 since they directly test stability and nuisance entanglement; the remaining cards typically require additional independent instruments or external outcomes beyond GoEmotions.

Card 1. We instantiate the recommended perturbations by swapping close encoders (all-MiniLM-L6-v2 vs all-MiniLM-L12-v2), varying aggregation (mean vs CLS pooling), and applying a simple style-normalization (original vs lowercased, punctuation-stripped text), yielding $k = 8$ variants (2 encoders \times 2 pooling rules \times 2 normalization conditions) on $n = 2000$ test examples. Each variant yields a reasonable proxy (AUC range 0.9407–0.9662). Treating variants as “raters,” absolute-agreement reliability is ICC(2,1)=0.8467 and ICC(2,k)=0.9779 (Shrout and Fleiss, 1979; Koo and Li, 2016).

Card 3. Let $\hat{C}_i = p(L_i = 1 \mid x_i)$ be a concrete proxy from sentence embeddings (Reimers and Gurevych, 2019). We form nuisance blocks Z from (i) length/style features (token/character length, “!”/“?” counts, uppercase ratio) and (ii) a topic block (TF-IDF + SVD). Discriminant regression (Eq. (6)) shows \hat{C} is largely nuisance-recoverable ($R^2=0.0245$ from length/style only; 0.7762 from topic only; 0.7768 from full Z). Incremental prediction (Eq. (7)–(8)) improves from AUC 0.9658 ($Z \rightarrow L$) to 0.9831 ($Z + \hat{C} \rightarrow L$), with $\beta_{inc} > 0$.

anchors and errors (high/low/borderline). High- C : “Thanks!!”, “Thank you!”, “Thanks”; Low- C : “No I’m not”, “No no she was [NAME]”,

“Why? I love it.” Borderline: “Thank [NAME] for the person walking down the sidewalk. . .”, “Did you kill the shark. . . Thanks [NAME]!”; a salient FP is “Yeah thank you you ungrateful bast. . .”, and FNs include “That’s because [NAME] thinks he’s the man. . .” and “Daily would be great, but. . . work. . .”.

5 Forensic Analysis: The Jangle Fallacy in Practice

To assess how often social NLP work substantiates the leap from an embedding-based proxy to a theoretically valid measure, we conducted a forensic coding of 17 influential papers (2020–2025) in ACL/NAACL/EMNLP that propose computational metrics for abstract social constructs. Our goal is not to single out individual contributions, but to evaluate whether the literature, in aggregate, separates (i) an *empirical instrument* (a scoring function over text representations) from (ii) the *theoretical construct* it claims to measure. Following standard measurement theory, we coded each paper on seven dimensions of validity and identification: construct definition, face/content evidence, reliability/stability, convergent validity, discriminant validity, predictive/criterion validity, and handling of confounding covariates (see Appendix A for coding criteria).

Results. Table 1 summarizes the distribution of evidence. Two patterns stand out. First, many papers provide *conceptual framing* and basic plausibility checks: construct definition is frequently articulated (Construct Validity: 10 Yes, 7 Partial, 0 No), and most papers provide some face/content evidence (6 Yes, 11 Partial, 0 No), often in the form of annotation guidelines, curated examples, or qualitative sanity checks. Second, the evidence most diagnostic for *measurement identification* is comparatively scarce. Convergent validity is rarely established against an independent instrument at the same unit of analysis (1 Yes, 12 Partial, 4 No). Discriminant validity is never fully demonstrated in the strict sense of ruling out nuisance surrogacy (0 Yes, 11 Partial, 6 No). Predictive/criterion validity under an external-outcome criterion is uncommon (1 Yes, 3 Partial, 13 No). Finally, none of the sampled papers uses causal identification methods to isolate the construct from confounds; confounding is handled primarily through heuristics or associational controls (0 Yes, 14 Partial, 3 No). Reliability/stability is the most consistently reported quan-

titative check (11 Yes, 4 Partial, 2 No), reflecting the routine reporting of inter-annotator agreement and, less commonly, robustness to perturbations.

Interpretation. This evidence profile is consistent with the paper’s central mechanism. When a measure is implemented as a geometric function over an entangled representation, three threats dominate: (i) the score may track nuisance structure such as topic, style, length, or prompt phrasing; (ii) the mapping from geometry to construct is underdetermined without validation (metric indeterminacy); and (iii) observed associations can be driven by confounding rather than construct variation. The forensic table suggests that current practice more often documents that a proxy is *defined* and *stable enough to compute* than that it *isolates* the intended construct.

More concretely, “Partial” evidence in the sample frequently takes forms that are compatible with the proxy presumption. Convergent validity is often operationalized as triangulation with related proxies, alignment with expectations, or comparisons to adjacent tasks—useful checks, but weaker than agreement with an independent validated instrument. Discriminant validity is commonly addressed through design constraints (e.g., topic balancing, prompt standardization) or qualitative discussion, rather than explicit tests showing the score is not explained by nuisance variables *with respect to nuisance confounds*. Confounding is typically handled via filtering, matching, or including covariates in predictive models, which controls associational but does not support identification without additional assumptions.

Implication. These gaps matter for cumulative science. If the literature rarely demonstrates discriminant validity and confound isolation, then two papers that both report a measure of “bias” or “ideology” can plausibly be tracking different mixtures of construct and nuisance variance. In that regime, results may appear to cumulate under a shared construct label while the underlying variables are not commensurate—a textbook setting for the *jangle fallacy* (Kelley, 1927). The forensic results therefore support the paper’s broader claim: downstream utility and qualitative plausibility are often treated as sufficient, even though the evidence most necessary to distinguish construct variance from method variance remains limited. This motivates the Construct Validity Protocol in Section 4 as a minimum reporting standard for embedding-based social mea-

Table 1: Validity evidence across 17 social measurement papers (2020–2025). “Yes” indicates rigorous adherence to measurement standards; “Partial” indicates heuristic or indirect checks; “No” indicates that the dimension is not substantively evaluated.

Dimension	YES	PARTIAL	NO
Construct Validity	10	7	0
Face/Content Validity	6	11	0
Reliability / Stability	11	4	2
Convergent Validity	1	12	4
Discriminant Validity	0	11	6
Predictive Validity	1	3	13
Handling Confounders	0	14	3

sures.

6 Alternative Views

Objection 1: “Post-hoc Correlation is Sufficient.”

A common counter-argument is that if a proxy correlates with a human label, disentanglement is unnecessary. We disagree. A high correlation coefficient (e.g., $r = 0.7$) can be driven entirely by a confounding variable Z . For example, a “Toxicity” classifier might correlate well with human labels simply because both humans and models flag AAVE dialect as toxic (Zhou et al., 2022). Without counterfactual validity (proving the score changes only when C changes, not just when Z changes), the metric is a biased estimator of the construct.

More broadly, learned-proxy methodology shows that post-hoc correlation is insufficient for testing causal theories with learned proxies (Knox et al., 2022).

Objection 2: “Large Scale Solves This.” Another view posits that sufficiently large models (e.g., GPT-4) implicitly understand the difference between concepts and nuisance variables. While LLMs are powerful generators, their embeddings remain entangled representations of the training distribution. As demonstrated by the oracle encoder fallacy, a perfect encoder preserves *all* information, including the noise. Scale improves the *fidelity* of the embedding, but it does not automatically perform the *causal abstraction* required to separate C from Z . Validity requires active methodological intervention, not just passive scaling.

7 Conclusion

The migration of NLP from engineering benchmarks to Computational Social Science demands a

parallel maturation in how we *measure*. We have argued that much of today’s construct-oriented work implicitly relies on “measurement by renaming”—treating convenient geometric heuristics (e.g., cosine distance in embedding space) as if they were identified measures of latent social constructs. By making the measurement problem explicit in a simple data-generating view, $D = G(C, Z)$, we show why this practice is fragile: unsupervised representations generally encode mixtures of the target construct (C) and nuisance attributes (Z) such as topic, style, venue, time, and authorship. In this setting, raw similarity metrics are not guaranteed to isolate C and can instead track pipeline-induced variation.

To move from plausible proxies to cumulative measurement, we propose the **Construct Validity Protocol (CVP)** as a community standard. The CVP operationalizes a full pipeline: (i) **conceptualization** with explicit construct boundaries and a domain of observables, (ii) **face/content validation** through exemplar design and expert audit, (iii) **instrument design** that targets construct-relevant text while controlling confounds, and (iv) a **Validity Suite** that reports reliability/stability, convergent evidence against independent labels, discriminant and *incremental* evidence beyond topic/style controls, known-groups separation, and predictive/criterion tests with falsification outcomes. Within this framework, we introduced **Counterfactual Neutralization**—using LLM-generated counterfactual rewrites to hold Z fixed while varying construct-relevant content—and complementary tools such as orthogonal projection when appropriate.

Our aim is not to prohibit proxies, but to make them *testable*. Accuracy is about hitting the target; construct validity is about ensuring the target is the right one. Embedding-based “social variables” should therefore be accompanied by transparent validity evidence—a compact **Validity Card** that documents design choices, stability checks, and the full set of diagnostics. Adopting such standards would make results comparable across papers, reduce the risk of topic/style leakage masquerading as social signal, and ultimately enable NLP-based measurement to serve as credible scientific evidence rather than convenient geometry.

Limitations

Our coding reflects what is *reported* in papers rather than what authors may have performed but

did not document. This is a feature—because cumulative science depends on transparent validity arguments—but it also means we may undercount validity evidence that exists only in unpublished analyses, code repositories, or informal checks. Moreover, our coarse {Yes/Partial/No} labels compress a spectrum of practices, and several dimensions (especially predictive/criterion validity and confound isolation) depend on definitional choices about what constitutes a sufficiently external criterion or an identification strategy. Future work could improve reliability by preregistering the rubric, double-coding with independent raters, and reporting inter-rater agreement.

Second, this paper is a position and synthesis contribution: we propose the Construct Validity Protocol (CVP) as a standard for measurement identification, but we do not instantiate the full protocol end-to-end on a new dataset. Empirical case studies that apply the CVP prospectively—including controlled discriminant tests and confound-neutralization designs—are needed to quantify the practical costs, failure modes, and benefits of the protocol in real measurement pipelines, especially as LLM-based encoders and prompting practices continue to evolve rapidly.

References

- Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3):529–546.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI magazine*, 36(1):15–24.
- Dilshod Azizov, Zain Muhammad Mujahid, Hilal AlQuabeh, Preslav Nakov, and Shangsong Liang. 2024. SAFARI: Cross-lingual bias and factuality detection in news media and news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12217–12231. Association for Computational Linguistics.
- Leif Azzopardi and Yashar Moshfeghi. 2025. POW: Political overtone windows of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24767–24773. Association for Computational Linguistics.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4982–4991.
- Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159.
- Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.
- Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.
- Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori B Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

- International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4040–4054.
- Robert F. DeVellis. 2016. *Scale Development: Theory and Applications*. Sage Publications.
- Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2022. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.
- Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36:68589–68601.
- Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.
- Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a little to the left: A theory-grounded measure of political bias in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31684–31704.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762.
- Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers’ reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, and 1 others. 2020. Evaluating models’ local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.
- R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 325–336.
- Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature*, 57(3):535–574.
- Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.
- Fabrice Y Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Senahan Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. Measuring psychological depth in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17162–17196.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, and 1 others. 2009. The elements of statistical learning.
- Stephen N Haynes, David Richard, and Edward S Kubany. 1995. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological assessment*, 7(3):238.
- Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, and 1 others. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.
- Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

- John Hunsley and Gregory J Meyer. 2003. The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. *Psychological assessment*, 15(4):446.
- Alexandra Ils, Dan Liu, Daniela Grunow, and Steffen Eger. 2021. Changes in european solidarity before and during covid-19: Evidence from a large crowd- and expert-annotated twitter dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1623–1637.
- Truman Lee Kelley. 1927. *Interpretation of educational measurements*. World Book Company.
- Dean Knox, Christopher Lucas, and Wendy K Tam Cho. 2022. Testing causal theories with learned proxies. *Annual Review of Political Science*, 25:419–441.
- Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.
- Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.
- Charles H Lawshe. 1975. A quantitative approach to content validity. *Personnel psychology*, 28(4).
- Junha Lee, Jaeshin Cho, Youngjin Cho, Hyewon Jin, Hyemin Lee, and Min Song. 2024. [ICScore: Metrics for evaluating interestingness and creativity of stories](#). Poster, NeurIPS 2024 Workshop: Statistical Frontiers in LLMs and Foundation Models.
- Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. 2010. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.
- Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374.
- Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.
- Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.
- R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3428–3448.
- William Merrill, Noah A Smith, and Yanai Elazar. 2024. Evaluating n-gram novelty of language models using rusty-dawg. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14459–14473.
- Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9):741.
- J.C. Nunnally and I.H. Bernstein. 1994. *Psychometric Theory*. Number no. 972 in McGraw-Hill series in psychology. McGraw-Hill Companies, Incorporated.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the seventh joint conference on lexical and computational semantics*, pages 180–191.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7237–7256.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4902–4912.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 1668–1678.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malih Alikhani, and Junyi Jessy Li. 2022. Political ideology and polarization: A multi-dimensional approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 231–243.

CH-Wang Sky, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 1667–1682.

Anvesh Rao Vijjini, Rakesh R Menon, Jiayi Fu, Shashank Srivastava, and Snigdha Chaturvedi. 2024. Socialgaze: Improving the integration of human social norms in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16487–16506.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423.

A Measurement Validity Coding Scheme and Rubric

This appendix specifies the coding rubric used in the forensic analysis (Appendix B). Each paper is evaluated on seven dimensions of measurement validity and identification. For each dimension, we assign one of three labels: YES, PARTIAL, or NO.

A.1 General Scoring Labels

- **YES (High rigor).** The method meets premium social science standards. It is grounded in established theory and/or uses rigorous psychometric validation or explicit causal designs, with clear empirical evidence.
- **PARTIAL (Medium rigor).** The method relies on operational heuristics, “silver standards,” or ad-hoc sanity checks. It may acknowledge threats (e.g., confounding) but addresses them with standard supervised learning, informal qualitative inspection, or design constraints rather than explicit identification.
- **NO (Low rigor).** The authors provide no discussion, definition, or empirical validation for the dimension.

A.2 Dimensions of Validity

Dimension 1: Target Variable Definition (Construct Validity).

- **YES.** Cites a specific, pre-existing social science theory (e.g., Moral Foundations Theory) and defines the construct independently of the dataset and measurement procedure (e.g., “We measure ideology as conceptually defined by Converse (1964) . . .”).
- **PARTIAL.** Defines the variable operationally, tautologically, or based primarily on dataset labels or a single chosen proxy (e.g., “toxicity is whatever the Perspective API labels as toxic,” or “bias is cosine distance between vectors”).
- **NO.** No clear definition is provided; the variable is treated as self-explanatory.

Dimension 2: Face Validity (Content Validity).

- **YES.** Conducts a formal content validity study prior to deployment, such as expert-panel review of items/lexicon or a structured pilot with domain experts.
- **PARTIAL.** Provides informal sanity checks (e.g., selected high/low scoring examples, qualitative inspection of top features/words) without a structured expert evaluation.
- **NO.** No inspection of instrument content and no qualitative examples are provided.

Dimension 3: Reliability / Stability.

- **YES.** Reports formal, chance-corrected reliability/stability metrics. For human annotation: Cohen’s/Fleiss’ κ , ICC, or Krippendorff’s α . For model-based instruments: test–retest reliability or rigorous robustness checks (e.g., prompt perturbations/paraphrasing) with appropriate agreement metrics.
- **PARTIAL.** Reports weak or non-corrected metrics (e.g., raw % agreement without chance correction), or evaluates stability only on a small/non-representative subset.
- **NO.** No reliability metric or stability check is reported.

Dimension 4: Convergent Validity.

- **YES.** Demonstrates significant association with a completely independent, external “gold-standard” measure of the same construct at the appropriate unit of analysis (e.g., “Our ideology score correlates with DW-NOMINATE roll-call votes”).
- **PARTIAL.** Correlates with a “silver standard,” a related-but-distinct proxy, or internal metadata that is suggestive but not a gold-standard instrument (e.g., star ratings, hashtags, outlet categories).
- **NO.** No external correlation or comparison is reported.

Dimension 5: Discriminant Validity.

- **YES.** Explicitly tests and empirically demonstrates that the measure is distinct from nuisance factors (e.g., near-zero correlation with document length, explicit topic residualization, or formal tests against style/genre/demographic/prompt effects).
- **PARTIAL.** Acknowledges potential confounds and/or offers qualitative arguments or design constraints, but does not provide formal tests sufficient to rule out nuisance surrogacy.
- **NO.** No discussion or test of discriminant validity.

Dimension 6: Predictive Validity.

- **YES.** Predicts a real-world downstream outcome external to the annotation/task setting (e.g., “polarization predicts future protest violence,” “trust predicts trading volume”).
- **PARTIAL.** Predicts an internal proxy outcome or trivial metadata label (e.g., predicting subreddit labels, publication year, or other dataset-internal fields).
- **NO.** No predictive/criterion task is reported.

Dimension 7: Handling Confounding Covariates.

- **YES.** Uses explicit causal inference methods to isolate the construct from confounding covariates Z (e.g., instrumental variables, double machine learning, propensity score matching, or rigorous residualization under stated assumptions).
- **PARTIAL.** Uses heuristics to exclude/limit Z , or uses standard supervised learning that includes Z as a feature (associative “controls” without causal identification).
- **NO.** No discussion of confounding covariates; treats relationships as direct without controls.

B Paper-Level Forensic Coding Notes

This appendix reports the paper-level coding notes underlying Table 1 in the main text. We code 17 influential ACL/NAACL/EMNLP papers (2020–2025) that propose computational measures of abstract social constructs using seven dimensions: (D1) target-variable definition (construct validity), (D2) face/content validity, (D3) reliability/stability, (D4) convergent validity, (D5) discriminant validity, (D6) predictive/criterion validity, and (D7) handling confounding covariates. Each dimension is coded as YES, PARTIAL, or NO following the rubric in Appendix A.

B.1 Paper-Level Coding Notes (D1–D7)

Paper 1: Demszky et al. (ACL 2020) — Fine-grained Emotions (Reddit).

- **D1 (Construct): PARTIAL.** Label set motivated by psychology literature and a careful selection process, but the construct definition remains closely tied to the taxonomy rather than an independent theory-grounded construct map.

- **D2 (Face): PARTIAL.** Presents labeled examples (e.g., example annotations), but no formal expert content-validity audit is reported.
- **D3 (Reliability): YES.** Reports rater agreement metrics including chance-corrected statistics (e.g., Cohen’s κ) alongside corroborating checks.
- **D4 (Convergent): PARTIAL.** Shows transfer/generalization to existing emotion benchmarks; useful triangulation but not a classic external gold-standard convergence test.
- **D5 (Discriminant): NO.** No explicit tests that emotion labels are distinct from nuisance factors (topic/style/length).
- **D6 (Predictive): NO.** Validations focus on label prediction rather than external real-world criteria.
- **D7 (Confounding): PARTIAL.** Identifies confounds and applies curation measures, but no causal identification strategy is used.

Paper 2: Baly et al. (EMNLP 2020) — Political Bias and Factuality (Outlets + Social Profiles).

- **D1 (Construct): PARTIAL.** Targets defined primarily through MBFC label categories rather than an independent theory definition.
- **D2 (Face): PARTIAL.** Informal sanity checks (e.g., excluding ill-defined label categories), but no structured content-validity study.
- **D3 (Reliability): NO.** Reliability of the inherited labels is not quantified in the paper.
- **D4 (Convergent): NO.** Evaluations mainly predict the same label scheme; no external convergence at the same unit of analysis.
- **D5 (Discriminant): NO.** Feature ablations are not framed as discriminant tests against nuisances.
- **D6 (Predictive): NO.** No external outcome prediction beyond the dataset labels.
- **D7 (Confounding): PARTIAL.** Practical heuristics (filters/exclusions/multi-source signals), but no causal identification methods.

Paper 3: IIs et al. (ACL 2021) — Social Solidarity / Anti-Solidarity (Twitter).

- **D1 (Construct): YES.** Provides an explicit, theory-grounded definition of solidarity.
- **D2 (Face): YES.** Involves social-science experts in refinement/adjudication.
- **D3 (Reliability): YES.** Reports Cohen’s κ and compares agreement across conditions.
- **D4 (Convergent): NO.** No convergence test against an external solidarity index/instrument.
- **D5 (Discriminant): PARTIAL.** Reports limited discriminant evidence (e.g., low correlation with sentiment), but no systematic nuisance-factor testing.
- **D6 (Predictive): PARTIAL.** Correlates the construct with external indicators (e.g., COVID-19 rates) at aggregate level; suggestive but not a dedicated predictive validation.
- **D7 (Confounding): PARTIAL.** Notes sampling and causal ambiguity; no identification strategy is implemented.

Paper 4: Demszky et al. (ACL 2021) — Conversational Uptake (Student–Teacher Transcripts).

- **D1 (Construct): YES.** Defines uptake as a linguistic/social construct and provides an operational definition.
- **D2 (Face): PARTIAL.** Offers qualitative plausibility checks and examples but no formal content-validity study.
- **D3 (Reliability): YES.** Reports structured annotation and inter-rater agreement with explicit statistics; describes aggregation.
- **D4 (Convergent): YES.** Tests alignment with indicators expected to track uptake.
- **D5 (Discriminant): PARTIAL.** Shows uptake is not reducible to repetition/overlap, but does not fully rule out broader nuisances.
- **D6 (Predictive): YES.** Links uptake to downstream outcomes relevant to teaching contexts (e.g., satisfaction/quality).
- **D7 (Confounding): PARTIAL.** Addresses some topical overlap concerns but does not implement causal identification.

Paper 5: Vidgen et al. (ACL 2021) — Online Hate (Adversarial Dataset).

- **D1 (Construct): YES.** Provides an explicit definition of hate for the annotation task.
- **D2 (Face): YES.** Emphasizes expert annotators and presents credibility signals for labeling.
- **D3 (Reliability): YES.** Reports chance-corrected agreement (e.g., Krippendorff’s α) across rounds.
- **D4 (Convergent): PARTIAL.** Evaluates against an external functional test suite (triangulation rather than gold-standard correlation).
- **D5 (Discriminant): PARTIAL.** Uses hard negatives/contrastive design to reduce keyword confounds, but does not provide formal discriminant statistics.
- **D6 (Predictive): NO.** Focuses on dataset/model evaluation, not external real-world outcomes.
- **D7 (Confounding): PARTIAL.** Mitigates confounds via data design, but no causal identification strategy is used.

Paper 6: ElSherief et al. (EMNLP 2021) — Latent Hatred / Implicit Hate (Twitter).

- **D1 (Construct): YES.** Develops a typology targeting implicit hate beyond explicit slurs.
- **D2 (Face): PARTIAL.** Taxonomy and examples support plausibility, but no formal expert panel content-validity study is reported.
- **D3 (Reliability): YES.** Reports chance-corrected reliability (e.g., ICC, Fleiss’ κ).
- **D4 (Convergent): PARTIAL.** Compares against widely used systems; informative but not an external gold-standard construct instrument.
- **D5 (Discriminant): PARTIAL.** Distinguishes implicit from explicit hate; limited evidence against adjacent nuisances (offensiveness/negativity).
- **D6 (Predictive): NO.** Evaluations are mainly benchmarking/model performance.

- **D7 (Confounding): PARTIAL.** Uses heuristics (e.g., filtering/keyword-related controls) to focus on the implicit construct, but not causal identification.

Paper 7: Liu et al. (NAACL 2022) — Ideology and Stance (News; POLITICS pretraining).

- **D1 (Construct): PARTIAL.** Ideology operationalized via outlet-level labels rather than theory-grounded definition.
- **D2 (Face): PARTIAL.** Informal plausibility checks (e.g., attention visualization), but no structured content-validity procedure.
- **D3 (Reliability): NO.** No reliability metrics reported for outlet-level ideology labels.
- **D4 (Convergent): PARTIAL.** Uses third-party outlet ratings and tests downstream utility, but no independent political-science ideology measure at same unit.
- **D5 (Discriminant): PARTIAL.** Limited topic-control evidence; no systematic nuisance testing.
- **D6 (Predictive): NO.** Focuses on ideology prediction/benchmarks rather than external outcomes.
- **D7 (Confounding): PARTIAL.** Uses heuristics to reduce imbalance/bias; no causal identification methods.

Paper 8: Sinno et al. (NAACL 2022) — Multi-dimensional Political Ideology (News Paragraphs).

- **D1 (Construct): YES.** Political ideology defined with political-science grounding and operationalized across dimensions.
- **D2 (Face): YES.** Provides interpretable examples and expert annotation credibility signals.
- **D3 (Reliability): YES.** Reports chance-corrected agreement (Krippendorff’s α).
- **D4 (Convergent): PARTIAL.** Triangulates with outlet-bias signals; not a direct external instrument correlation.
- **D5 (Discriminant): PARTIAL.** Separates ideology from stance by design; no formal nuisance-confound tests beyond constraints.

- **D6 (Predictive): NO.** Evaluations are classification/label prediction.
- **D7 (Confounding): PARTIAL.** Uses design/annotation constraints (e.g., controlling for stance), but no verified causal isolation.

Paper 9: Gabriel et al. (ACL 2022) — Misinfo Reaction Frames (Headlines).

- **D1 (Construct): PARTIAL.** Defines dimensions operationally; construct remains task-specific rather than externally standardized.
- **D2 (Face): PARTIAL.** Provides examples/descriptions; crowd annotations without expert-panel content validity.
- **D3 (Reliability): PARTIAL.** Reports reliability for categorical judgment; limited/no parallel reliability reporting for free-text dimensions.
- **D4 (Convergent): NO.** No validation against external behavioral ground truth of reactions.
- **D5 (Discriminant): NO.** No explicit tests distinguishing the construct from nuisances.
- **D6 (Predictive): PARTIAL.** Shows systematic shifts in trust ratings in a controlled setting; not real-world behavioral criterion validation.
- **D7 (Confounding): PARTIAL.** Design mitigations exist, but no causal confound-handling pipeline.

Paper 10: Hoover et al. (2020) — Moral Sentiment (MFT; Tweets).

- **D1 (Construct): YES.** Categories grounded in Moral Foundations Theory.
- **D2 (Face): YES.** Provides structured guidelines and annotator training procedures; early disagreement handling.
- **D3 (Reliability): YES.** Reports chance-corrected agreement (e.g., Fleiss' κ , PABAK).
- **D4 (Convergent): PARTIAL.** No correlation with an independent external MFT instrument; convergence mostly indirect.
- **D5 (Discriminant): PARTIAL.** Conceptual separation of categories without formal nuisance-factor tests.

- **D6 (Predictive): NO.** Benchmarks classifiers; no external real-world criterion validation.
- **D7 (Confounding): NO.** No confound-isolation or causal identification methods.

Paper 11: Feng et al. (ACL 2023) — Political Bias from Partisan Pretraining (PCT axes).

- **D1 (Construct): PARTIAL.** Operationalizes leaning via PCT outputs; construct remains tied to instrument choice rather than independent definition.
- **D2 (Face): PARTIAL.** Qualitative plausibility checks via examples; no formal expert content-validity study.
- **D3 (Reliability): YES.** Reports chance-corrected agreement for a stance detector and includes prompt-robustness analyses.
- **D4 (Convergent): PARTIAL.** Compares against media-bias ratings of pretraining sources (external reference point).
- **D5 (Discriminant): NO.** No explicit tests separating ideology from superficial keyword/prompt effects.
- **D6 (Predictive): PARTIAL.** Links pretraining bias to downstream fairness shifts on benchmark tasks; still not an external real-world criterion.
- **D7 (Confounding): PARTIAL.** Uses controlled model-building (e.g., comparable corpora sizes) to isolate the pretraining factor; not causal identification.

Paper 12: Sky et al. (EMNLP 2023) — Social Norms (Cross-cultural).

- **D1 (Construct): YES.** Defines norms explicitly and situates them in cross-cultural/descriptive-norm framing.
- **D2 (Face): YES.** Expert verification/editing plus concrete examples.
- **D3 (Reliability): PARTIAL.** Some chance-corrected agreement is reported, but not as a core-label inter-annotator reliability statistic.
- **D4 (Convergent): PARTIAL.** Aligns with established cross-cultural theory patterns; no independent gold-standard norm instrument correlation.

- **D5 (Discriminant): NO.** No explicit empirical separation from nuisance factors.
- **D6 (Predictive): NO.** No external outcome prediction beyond the dataset/task.
- **D7 (Confounding): NO.** No causal/confound-isolation methods.

Paper 13: Bang et al. (ACL 2024) — Political Bias as Stance + Framing/Style (LLM generations).

- **D1 (Construct): PARTIAL.** Bias defined operationally as stance plus framing/style rather than theory-grounded construct definition.
- **D2 (Face): PARTIAL.** Qualitative examples/sanity checks; no formal content-validity study.
- **D3 (Reliability): PARTIAL.** Uses repeated generations and significance testing, but does not report chance-corrected reliability of the measurement procedure.
- **D4 (Convergent): PARTIAL.** Compares against prompted anchor distributions rather than independent external instrument.
- **D5 (Discriminant): PARTIAL.** Separates content from lexical polarity but lacks formal nuisance-confound tests.
- **D6 (Predictive): NO.** Descriptive audit; no external criterion prediction.
- **D7 (Confounding): PARTIAL.** Standardizes prompts/topics and uses anchor distributions; not causal identification.

Paper 14: Harel-Canada et al. (EMNLP 2024) — Psychological Depth Scale (Stories).

- **D1 (Construct): YES.** Construct grounded in literary/reader-response theory and defined independent of dataset.
- **D2 (Face): PARTIAL.** Guidelines/training/calibration, but no formal expert content-validity audit.
- **D3 (Reliability): YES.** Reports chance-corrected inter-rater reliability for human ratings (e.g., Krippendorff’s α).

- **D4 (Convergent): PARTIAL.** LLM-judge scores correlate with human judgments, but not with an independent external gold-standard instrument.
- **D5 (Discriminant): PARTIAL.** Argues beyond surface style; lacks formal nuisance-confound tests.
- **D6 (Predictive): NO.** Used as evaluation rubric; no external real-world outcome prediction.
- **D7 (Confounding): PARTIAL.** Uses design controls; no causal identification.

Paper 15: Azizov et al. (EMNLP Findings 2024) — Political Bias and Factuality (Cross-lingual; xMP).

- **D1 (Construct): PARTIAL.** Targets operationalized via existing rating schemes rather than theory-defined constructs.
- **D2 (Face): PARTIAL.** Relies on expert-provided outlet-level ratings and supplemental labels; no formal content-validity study for the resulting instrument.
- **D3 (Reliability): PARTIAL.** Reports cross-check alignment between article samples and outlet labels; informative but not full stability analysis.
- **D4 (Convergent): NO.** No clear external convergent validation reported.
- **D5 (Discriminant): NO.** No explicit tests against nuisance confounds.
- **D6 (Predictive): NO.** Benchmarks models to predict labels; no external outcome criterion.
- **D7 (Confounding): NO.** No causal/confound-isolation methods.

Paper 16: Faulborn et al. (ACL 2025) — Political Values Measurement (WVS/EVS items).

- **D1 (Construct): YES.** Builds measurement around validated survey instruments and argues against ad-hoc quizzes.
- **D2 (Face): YES.** Inherits content validity from decades of survey-item vetting.

- **D3 (Reliability): YES.** Tests prompt sensitivity and demonstrates instability; reported as a stability diagnostic.
- **D4 (Convergent): PARTIAL.** Motivates survey validity and population alignment, but does not provide a clean correlation-style convergent test as coded.
- **D5 (Discriminant): PARTIAL.** Diagnoses wording/prefix contamination (nuisance sensitivity), but does not report a classic discriminant test against pre-specified nuisance factors.
- **D6 (Predictive): NO.** Focus is measurement validity rather than external outcome prediction using the derived score.
- **D7 (Confounding): PARTIAL.** Uses design-based mitigation; not causal identification.

Paper 17: Azzopardi and Moshfeghi (EMNLP Findings 2025) — Political Overton Window (PRISM audit).

- **D1 (Construct): YES.** Operationalizes Overton Window by mapping espouse/neutral/refusal across a spectrum.
- **D2 (Face): PARTIAL.** Visualizations provide plausibility, but no formal expert content-validity study.
- **D3 (Reliability): YES.** Reports chance-corrected agreement for the assessment procedure.
- **D4 (Convergent): PARTIAL.** Interprets boundaries relative to expected alignment behavior; no independent ideology instrument correlation.
- **D5 (Discriminant): PARTIAL.** Distinguishes position from acceptability conceptually; limited empirical nuisance-separation evidence.
- **D6 (Predictive): NO.** Descriptive audit; no external criterion prediction.
- **D7 (Confounding): PARTIAL.** Persona probing and demographic/prompt heuristics; no causal identification methods.