

The Measurement Revolution? Credible Measurement and Inference in the Age of AI*

Melissa Dell and Ashesh Rambachan

August 2026

Abstract

Artificial intelligence (AI) is transforming measurement in economics. AI models convert unstructured data, such as text and images, into structured variables at low cost, making previously prohibitive measurement feasible at scale. This shifts the bottleneck from finding *any* scalable measure of a phenomenon to choosing among *many* plausible ones, which may support different empirical conclusions. This review provides guidance for navigating that shift. We describe three stages at which AI enters the measurement pipeline—discovery, construct definition, and observation—and what each demands of researchers. We argue that credible inference with AI-generated variables requires appropriately designed validation: anchoring measurement to explicit criteria, rather than informal claims that a proxy is reasonable. We then examine how validation samples support valid inference even when AI predictions are arbitrarily biased, and what can be done when a random validation sample is unavailable.

Keywords: measurement, artificial intelligence, unstructured data, measurement error.

*This review accompanies the National Bureau of Economic Research 2026 Methods Lecture on *Estimation and Inference with AI-Generated Data*, available at <https://www.nber.org/conferences/si-2026-methods-lecture-estimation-and-inference-ai-generated-data>.

1 Introduction

Artificial intelligence (AI) is transforming measurement in economics. AI models can convert unstructured data (*e.g.*, text, images, audio, and video) into low-dimensional structured variables for economic analysis at very low marginal cost. For example, large language models (LLMs) can code open-ended survey responses, historical documents, and massive-scale web text corpora, while computer vision models can transform satellite imagery into measures of poverty, deforestation, and pollution at fine spatial resolution and with broad geographic coverage. AI makes measurement endeavors that were once prohibitively costly feasible at scale, allowing economists to revisit long-standing questions with new evidence and to study questions that were previously out of reach.

Although measurement is often conceived as passive observation through an instrument—whether a ruler or an LLM—in economics it is better understood as a disciplined simplification of a complex, high-dimensional world. Jorge Luis [Borges \(1998\)](#) vividly captures this in “On Exactitude in Science,” where he imagines cartographers who pursue perfect representation until they produce a map “whose size was that of the Empire.” The perfect map, however, is useless: because it no longer reduces complexity, it ceases to function as a map. Our goal is not to produce a perfect measure of economic reality, but to construct a simplified representation that reveals the broader landscape and helps navigate it.

More formally, measurement specifies and implements a function $m(\cdot)$ that projects high-dimensional reality U into a lower-dimensional representation $m(U)$ that can be interpreted and analyzed. For example, economists have mapped newspaper articles (high-dimensional texts) into an indicator for whether the article discusses economic policy uncertainty ([Baker, Bloom and Davis, 2016](#)). Because U is high-dimensional, there are often many possible ways to measure the same concept. Different measurement functions $m_j(\cdot)$ may preserve different features of U and may therefore support different empirical conclusions.

Deep neural networks are the foundation of modern AI, and one of their most common uses is as flexible implementations of measurement functions (for a survey aimed at economists, see [Dell, 2025](#)). Earlier approaches to unstructured data often encoded human judgment directly through hand-engineered rules. Neural networks instead learn from empirical examples, adjusting millions or billions of parameters to perform a specified task, such as predicting the next word or classifying an image. Given sufficient training data and an appropriate training objective, they learn parameters that compress U into a lower-dimensional representation to perform the task at hand, without requiring the researcher to specify which features should be preserved, discarded, or transformed. The representations they learn can often be reused for other tasks, a principle known as “transfer learning.”

Because neural networks are highly scalable, they can dramatically lower the marginal cost of implementing measurement functions $m(\cdot)$, extending systematic measurement to domains that were previously out of reach. The size of the overall cost reduction depends on the task. In some settings, AI measurement requires substantial fixed investments: to achieve sufficiently accurate predictions, researchers need to develop high-quality training data and fine-tune customized models, so constructing $m(\cdot)$ remains costly although applying it at scale is cheap. In other settings, off-the-shelf or lightly adapted models, such as commercial LLMs, perform sufficiently well to dramatically lower both fixed and marginal costs.

Cheap measurement makes it feasible to construct many alternative measures of the same underlying concept, and AI introduces many implementation choices that can meaningfully affect the resulting measure. Researchers can vary the measurement rubric, model, and prompt; when fine-tuning, they must also choose the training-data distribution and tuning strategy. Model parameters add further degrees of freedom, even when defaults are available.

This abundance of potential measures marks a substantial shift in empirical work. Researchers have often had to rely on externally constructed measures as fixed proxies for theoretically meaningful concepts—for example, using an external measure of expropriation risk for foreigners as a proxy for property rights institutions. Such proxies may contain arbitrary measurement error or poorly align with the construct of interest, but they were often the only scalable option. With AI, the bottleneck is shifting from finding *any* scalable measure of a phenomenon to choosing among *many* plausible ones.

Interpreting differences across abundant AI measures is complicated. While extremely powerful, deep neural networks are in many ways black boxes. When measurement is delegated to an AI model, the features that are preserved in $m(U)$ are determined implicitly through complex interactions among millions or billions of learned parameters. Direct inspection of these parameters generally does not reveal an interpretable understanding of how the model constructs $m(U)$. This opacity may not matter when the end goal is prediction (Mullainathan and Spiess, 2017), but it becomes consequential when AI-generated variables are used in downstream empirical analysis (Ludwig, Mullainathan and Rambachan, 2026). Economists seek to interpret relationships involving $m(U)$, which requires understanding what measures capture and why alternative measures of the same concept can yield different conclusions, a phenomenon documented by a growing literature reviewed in this article.

These considerations translate into several concrete challenges for incorporating AI-generated measures into empirical analysis. The researcher must choose which measure(s) to use and report, and post-selection inference concerns can arise (e.g., Baumann et al., 2025). This is compounded by the fact that researchers can generate large volumes of noisy, poorly validated, or difficult-to-interpret measures, what is colloquially called “AI slop.”

Researchers must decide when costly investments in training data, fine-tuning, or model refinement are worthwhile to improve poorly performing measures. This is difficult to assess without a principled framework for how measurement quality affects estimates of the target parameter. Finally, reproducibility concerns arise because proprietary models are frequently deprecated (e.g., [Barrie, Palmer and Spirling, 2024](#); [Coqueret et al., 2026](#)).

How should economists approach AI measurement given these challenges? To elucidate this question, this review first describes the three stages at which AI can enter the measurement pipeline, and what each stage demands of the researcher (Section 2). *Discovery* selects which features of the world to measure at all. *Construct definition* fixes what the selected concept means, in terms precise enough to apply to a specific setting. *Observation* implements that definition at scale, producing the variable that enters the analysis. AI can propose candidate features directly from the data, turn a single concept into many competing operationalizations, and apply a definition to millions of units at minimal marginal cost.

The existence of many potential measures does not make measurement inherently subjective. Rather, it underscores the importance of validating whether measurements satisfy clearly articulated objectives, which are themselves subject to scientific debate. Credible inference with AI-generated predictions requires appropriately designed measurement validation. We define validation as anchoring measurement to explicit, observable criteria, rather than simply relying on informal arguments that a proxy is reasonable, and discuss the evidence for its importance (Section 3). Diverse literatures from econometrics, statistics, biostatistics, machine learning, psychometrics, and computational social science provide a well-developed foundation for designing validation, while also highlighting important gaps in existing methods.

Statistical methods are the most developed for the observation stage of measurement, and this is where AI is making the most rapid inroads. Hence, we largely focus on observation in this review. We examine how validation samples can support valid inference even when AI predictions are arbitrarily biased (Section 4). The key is that prediction errors are learned from the validation design, rather than inferred from assumptions about how the model generates its outputs. A random validation sample, however, is not always available. Credibility then rests on assumptions about how AI models behave across settings, rather than on the design of the validation sample (Section 5). Those assumptions can be stated precisely, and some can be checked against evidence on how frontier models err. However, this remains a largely open frontier given the enormous complexity of deep neural networks.

The opportunities created by AI, and the credibility challenges that accompany them, echo an earlier transformation in economics. When personal computing made estimation cheap in the 1990s, researchers could run not just one empirical specification but hundreds,

thousands, or even millions (Sala-I-Martin, 1997). That forced the field to ask what should discipline the choice among many possible specifications. The response was the credibility revolution (Angrist and Pischke, 2010; Angrist, 2022): rich literatures developed around causal inference, robustness, and multiple hypothesis testing, while empirical practice increasingly emphasized specifications designed to support clear interpretation.

AI moves specification search upstream, from choosing among empirical specifications to choosing among alternative measurements of a high-dimensional reality. The relevant lesson is analogous: expanding measurement choice highlights credibility problems that require empirical discipline and transparency to address. Researchers should specify verifiable measurement aims, evaluate how well their measurements achieve those aims, and design measurement to support valid, interpretable inference. This article synthesizes both classic and emerging literatures that build the foundation for credible measurement in the AI era.

2 The Measurement Pipeline: Discovery, Definition, and Observation

Colloquially, measurement is often viewed as the act of observing a quantity with an instrument or procedure. For economic research, it is more useful to view measurement as a pipeline with three stages. First, *discovery*: which features of the world should the researcher measure? Second, *construct definition*: how should the relevant concept be operationally defined, given the research question? Third, *observation*: how can that definition be implemented at scale? A researcher might, for example, identify economic policy uncertainty as the feature of interest, write a detailed rubric specifying what counts as economic policy uncertainty and what does not, and implement that rubric by prompting an LLM or designing a keyword query over economic news articles (Baker, Bloom and Davis, 2016).¹

Each stage carries a different kind of uncertainty, connects to different methodological literatures, and calls for a different form of validation. Discovery involves uncertainty about the structure of the world: which patterns, dimensions, or relationships exist and are worth measuring? It has historically been guided by theory and exploratory analysis, and it requires evidence that a procedure has uncovered meaningful structure. Construct definition involves conceptual uncertainty: how should a particular phenomenon be translated into an operational construct? It draws again on theory and domain expertise but has also spurred

¹Statistical inference is closely related: it maps structured data into a lower-dimensional representation, such as a target parameter, and could itself be viewed as a stage of the measurement process. We reserve measurement for the construction of variables from underlying reality, and inference for the use of those variables to learn about a target parameter.

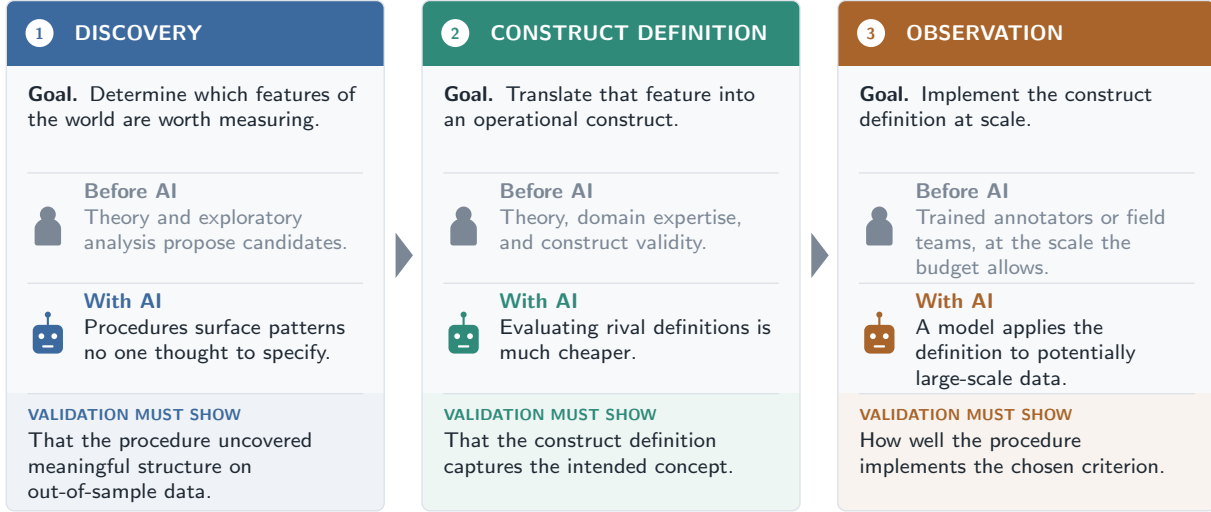


Figure 1: **The measurement pipeline.** Measurement proceeds in three stages: discovery, construct definition, and observation. AI is transforming each stage, creating both opportunities and challenges for empirical research.

construct validity literatures in psychometrics and survey design. It requires evidence that an operationalization captures the intended concept. Finally, observation involves statistical uncertainty from imperfect implementation: how accurately can a construct be measured at scale? It draws on the measurement error and semiparametric inference literatures, and it requires evidence that a procedure implements the chosen criterion.

AI is transforming all three stages of the measurement pipeline. In discovery, it can surface patterns that no one thought to specify in advance. In construct definition, it can lower the cost of generating and comparing alternative operationalizations of a concept. Finally, in observation, it can apply a chosen definition at scale. Figure 1 summarizes the three stages, what each requires, and how AI is entering each of them. This section introduces each stage in turn. The remainder of the article then concentrates on observation, where the methods for working with AI-generated measurements are the most developed, and AI use is accelerating most rapidly.

2.1 Discovery

Discovery has typically been guided by theory, contextual knowledge, or incremental advances in an existing literature. This is productive when a problem is well enough understood that pre-specified measurement is warranted. But it also means that the structured datasets anchoring empirical economics contain only the features that someone previously

decided to measure: the closed survey items we field, the outcomes we track in a randomized experiment, the variables of an administrative extract. Unstructured data record much richer information. Relying on researcher intuition and creativity to decide what to measure in these data can produce a bottleneck, slowing what we can discover.

In data-driven discovery, the researcher specifies a procedure rather than a predefined construct. When the researcher has a particular outcome in mind, these procedures can be directed toward uncovering constructs that relate to it. This activity is commonly known as “hypothesis generation.” The resulting structure enters measurement as an input to construct definition, where it is refined into a concept and then evaluated in a separate sample, or it may itself serve as the target measurement.

These procedures are not hypothetical: across modalities, they have already produced validated discoveries. [Ludwig and Mullainathan \(2024\)](#) turn the camera on the judge, predicting pretrial detention decisions from defendants’ mug shots. Much of the mug shot’s predictive signal is unexplained by structured characteristics or known facial features: something unnamed in the image systematically drives detention decisions. Their hypothesis generation procedure names two novel hypotheses, well-groomed and heavy-faced, that survive validation on held-out data. [Baron et al. \(2026\)](#) bring the same logic to text, using the rich case notes written by child protective services investigators to generate hypotheses about what separates high-performing investigators from the rest. [Obermeyer et al. \(2026\)](#) extend it to medical waveforms, morphing electrocardiograms along a clinical predictor to surface a novel feature of the waveform that predicts sudden cardiac death.

The broader literature on hypothesis generation is rapidly growing, spanning economics, computer science, and beyond (see [Mullainathan and Rambachan, 2025](#), for a brief overview). Applications now include effective teaching ([Workman, 2025](#)), medical notes ([Donahue et al., 2026](#)), and cognitive experiments ([Zhu et al., 2026](#)). Hypothesis generation techniques also enable researchers to collect and then analyze open-ended survey responses: AI-conducted interviews make open-ended elicitation cheap at scale ([Haaland et al., 2024](#); [Chopra and Haaland, 2026](#); [Geiecke and Jaravel, 2026](#)), and hypothesis generation procedures can automate the discovery of recurring themes in the resulting responses ([Wang, Saperstein and Pierson, 2026](#)). An accompanying methodological literature on hypothesis generation studies how to describe the differences between text corpora ([Zhong et al., 2022, 2023](#)), generate hypotheses with large language models ([Zhou et al., 2024](#); [Batista and Ross, 2024](#)) and sparse autoencoders ([Movva et al., 2025, 2026](#); [Peng et al., 2026](#)), conduct statistical inference on generated hypotheses ([Modarressi, Spiess and Venugopal, 2025](#); [Carlson, 2026](#)), and benchmark hypothesis generation procedures ([Liu et al., 2025](#)).

Discovery is disciplined not by how hypotheses are generated but by how they are evalu-

ated. Evidence on the quality of a generated hypothesis must come from held-out data: the researcher measures the named construct there and asks how well it predicts the outcome of interest. If the generating procedure never touches that held-out sample, it does not matter how the hypothesis was produced, and procedures as different as an expert’s intuition and a large language model can be judged by the same standard. This is the logic of the “common task framework,” which has organized much of the progress in modern machine learning (Donoho, 2024).

Nonetheless, a generated hypothesis is only the beginning. Discovery yields a named concept. To study that concept, for example estimate its prevalence or include it in a regression, the researcher must define it precisely and measure it across the full dataset. Those are the tasks of the next two stages.

2.2 Construct Definition

Construct definition asks how a given concept should be operationalized in a particular research setting. This stage is difficult because many concepts of interest are multidimensional, not directly observable, and closely relate to other phenomena. A measure of trust, for example, should capture the dimension of trust relevant to the research question without simply reproducing institutional quality, income, or other correlated variables. Theory and domain expertise therefore remain central: they supply the criteria against which an operationalization can be judged.

AI makes it cheap to generate operationalizations themselves. A researcher can ask a model to draft many rubrics for a concept like economic policy uncertainty, each plausible, each drawing the concept’s boundary somewhere slightly different. The researcher is then choosing among constructs, not merely among ways of implementing one.

Adjudicating among such candidates is an old problem, and the classic treatment is Cronbach and Meehl (1955). Many social science concepts, such as trust, social capital, and intelligence, are latent and complex. Since they cannot be directly observed, Cronbach and Meehl argue that validation must proceed *indirectly* through a “nomological network”: theory or contextual knowledge that implies what the construct should correlate with (referred to as convergent validity), what it should be distinct from (referred to as discriminant validity), and how stable it should be across settings. A construct is validated not by any single observable variable but by a broader pattern of relationships. In the language of econometrics, this logic is analogous to partial identification under a set of theoretically motivated moment restrictions: no single restriction establishes that an operationalization captures the intended construct, but additional restrictions narrow the set of plausible measures.

Construct validity is increasingly important for economics. The discipline has historically emphasized measures that are directly observed or for which direct validation is plausible, such as prices, schooling, and employment, devoting less attention to construct validity than work in psychometrics or survey design. Where such constructs are well established, construct validity is largely settled. But economists increasingly study complex latent concepts such as economic policy uncertainty, trust, and institutional quality, where construct validity is far more contested.

The abundance of candidate operationalizations is also an opportunity. When many are available, the researcher must say what distinguishes them, which forces the construct and its nomological network to be stated precisely enough to adjudicate among candidates. Where off-the-shelf models perform well enough, AI lowers the cost of that adjudication too, since the observable implications of the network, such as related outcomes, correlates, contrasts, and contextual patterns, can themselves be measured cheaply. What binds is not the supply of candidate measures but the theory and domain knowledge needed to say precisely what we are trying to measure. AI can draft rubrics and measure their implications; it cannot tell the researcher which operationalization answers the research question. That judgment rests on theory-driven measurement and substantive expertise about the setting. If the theoretical framework is weak, a measure's failure to align with the network may indict the underlying theory rather than the measure.

When credible external criteria are unavailable altogether, it may be better to focus on the discovery stage. Data-driven discovery can surface interpretable patterns without requiring strong priors about constructs that are not yet well understood. Those discovered patterns can then inform theory and construct definition.

2.3 Observation

In *observation*, AI serves as a scalable measurement technology. Once a researcher has specified an operational definition, an AI system can apply it repeatedly at low marginal cost, for example classifying large text corpora or extracting information from vast collections of images at a scale that would be infeasible for human coders.

Where construct validity asks whether an operationalization captures the concept needed for the research question, observational validity asks whether a procedure implements that operationalization. To understand this difference, a useful, though imperfect, analogy is the distinction between internal and external validity: observational validity (like internal validity) assesses credibility conditional on a specified target, while construct validity (like external validity) asks whether that target answers the broader question motivating the

research. The criterion may itself be a flawed operationalization of the concept of interest, but it is explicit and interpretable. Observation takes that criterion as fixed and asks only whether a procedure implements it faithfully.

Errors from AI-generated measurements are unlikely to be classical. Systematic biases can arise from network architecture, the distribution of training data, and other implementation details, and the nonlinear transformations applied at each layer of the neural network, together with the frequent use of binary or multiclass outputs, violate classical measurement error assumptions. These biases do not stay in the measurement; they propagate to estimates of the target parameter. AI-generated variables therefore cannot simply be substituted for the constructs they are meant to capture.

The remainder of the article takes up what to do instead: Section 3 asks what credible measurement with AI-generated variables requires, and Sections 4 and 5 give the two answers, depending on whether the researcher can construct a random validation sample.

3 From AI Outputs to Credible Measurements

3.1 Validation as an Anchor for Credible Measurement

Intuitively, validation plays an important role in establishing credibility, but what should be validated? Validation cannot show that a constructed measure captures a latent concept perfectly or recovers “fundamental truth.” Measurement aims at simplified representations of a complex reality that preserve the features needed to draw useful inferences. Ideally, validation would therefore establish that a measure achieves that broader aim. This is often extremely challenging to assess quantitatively, because the underlying reality is complex, partially unobservable, multi-causal, and not easy to manipulate experimentally.

When a measure is treated as a proxy, it is common to argue (often qualitatively) why it achieves this aim. But qualitative arguments often cannot adjudicate among many plausible measures that could lead to different conclusions. Consider economic policy uncertainty. One researcher counts newspaper articles containing terms for the economy, policy, and uncertainty; another asks a language model whether an article conveys uncertainty about economic policy; a third restricts attention to federal policy alone. Each choice can be defended as a reasonable proxy, and each may produce a different series.

Validation therefore has a simpler—and more achievable—empirical aim. It requires specifying observable measurement criteria, assessing how candidate measures behave relative to those criteria, and adjusting target estimates for systematic errors in measurement. In the economic policy uncertainty example, it means committing to a rubric—supported

by theory and domain expertise—that states what counts as expressing economic policy uncertainty and what edge cases do not. Competing measures can then be scored against that rubric, disagreements traced to particular articles, and systematic error corrected. Such criteria are themselves imperfect, but they can be stated explicitly and refined. This encourages scientific debate and incremental improvements in measurement for difficult-to-measure concepts.

Anchoring measurement to *explicit criteria* is especially important in the age of AI, because these criteria address a central limitation of modern AI systems. These systems are built from neural networks whose mappings from inputs to outputs are too complex to understand by inspecting their parameters, and whose learned criteria are not expressed as explicit, researcher-auditable rules. Their biases can shift across inputs in ways that humans predict poorly (Vafa, Rambachan and Mullainathan, 2024). Validation makes the black box of modern AI interpretable, not by elucidating its internal workings, but by evaluating its outputs against criteria specified by the researcher.

Validation also addresses the abundance of measures that AI can cheaply produce. When many candidate measures can be constructed, measurement choices themselves become objects of empirical comparison. Validation disciplines that comparison: it supplies the evidence for choosing among candidates, correcting or refining them, and deciding when further investment in measurement is warranted.

Validation does not privilege (fallible) humans over (fallible) AI. Rather, it privileges explicit, well-defined criteria as a way to make predictions from black-box models more interpretable. Validation evidence may come from human-coded labels, scientific instruments, administrative data, or theoretical predictions. Human codings are useful insofar as they are auditable implementations of a criterion.

Ambiguous edge cases and residual disagreement between annotations, while common, do not imply that validation lacks value. Instead, they identify where refinement may be needed: the construct definition may be ambiguous, the measurement procedure may be underspecified, or the validation criterion may otherwise require clarification.² A black-box neural network will often encode richer information than any given criterion. The question is whether this additional variation suggests a better construct, or instead reflects irrelevant features, or patterns that are difficult to interpret.

²Recent work builds AI-assisted pipelines for surfacing such edge cases (Xiong et al., 2025) and prompting the researcher to adjudicate them—another example illustrating how AI can transform the construct definition stage of measurement.

3.2 Which Model? Which Prompt?

It is useful to make the decision of whether and how to validate AI measures more concrete. Consider a researcher with a corpus of documents and a concept they would like to measure. The researcher constructs a careful rubric for the concept, converts it into a prompt, queries a frontier language model, and obtains a label for every document in the corpus at minimal cost. Rather than training annotators to score economic policy uncertainty in newspaper articles (Baker, Bloom and Davis, 2016) or the tone of congressional speeches about immigration (Card et al., 2022), the researcher simply asks the model. The temptation is to treat these outputs as the (noiseless) measurements themselves or proxies, and proceed directly to the downstream analysis. Is this a credible exercise?

A core challenge that such an approach must confront is that a well-defined construct alone does not pin down the measurement. The researcher must also specify a model, a prompt (when generative AI is used), and various model settings (that are often set by default but can sometimes be changed)—such as the temperature, decoding algorithm (*e.g.*, greedy decoding), and system prompt. The prompt engineering strategy determines how the rubric is converted into instructions for the model: as a system prompt or a user prompt; in full or in summary; with or without worked examples. The researcher may even ask the model to adopt a persona, such as an expert annotator. Each combination of choices produces an alternative measurement of the same concept, and generating yet another alternative is nearly free. Alternatively, if the researcher is not satisfied with off-the-shelf models, they may choose to train their own model, choosing an architecture, training data distribution, and a variety of other important details like the hyperparameters, the tokenizer, and pre- and post-processing steps. Such decisions are free parameters of the measurement process: nothing in the construct itself pins down the choice between AI models or prompt strategies.

These choices can matter enormously for downstream estimates. Varying the language model and the prompt engineering strategy can move the resulting estimates in magnitude, in statistical significance, and even in sign. Ludwig, Mullainathan and Rambachan (2026) document this sensitivity for measurements of congressional bills and financial news headlines, and a growing literature documents the same phenomenon across annotation tasks in the social sciences (Sclar et al., 2024; Atreja et al., 2025; Yin, Vu and Persico, 2026). Baumann et al. (2025) replicate annotation tasks from 21 published studies and show that, with a handful of prompt paraphrases, virtually any hypothesis can be made to appear statistically significant.³

³Recent work builds structured prompting pipelines that make outputs more reliable and reproducible across semantically similar prompts (Asirvatham, Moksiki and Shleifer, 2026). This work is valuable and can potentially narrow the arbitrariness of these choices.

In the face of this evidence, one response simply accepts the sensitivity. If different protocols produce different measurements, perhaps we should define the concept to be the output of one particular model under one particular prompt and collection of model settings. This position is internally consistent, but its implications are hard to accept. Are researchers using ChatGPT, Claude, DeepSeek, or any other of a multitude of commercial, open-source, or customized models to measure economic policy uncertainty really studying different constructs? What about two researchers who use the same model but different prompts? And what are we to do when a new frontier model is released? Should we revisit every empirical finding based on the previous generation’s outputs? Defining the construct as the output of a specific AI model makes results incomparable across studies, even studies asking the same economic question. More fundamentally, the problem with this view is that measurement should be anchored in the construct, not the other way around.

A second response defends the choices by checking agreement. If several models or prompt strategies tend to assign similar labels, the reasoning goes, then the free parameters must not matter very much. However, agreement among AI-generated measurements is weaker evidence than it appears. Frontier models are built on similar architectures, trained on overlapping corpora, and tuned toward similar preferred answers. An emerging literature on response homogenization documents exactly this convergence (Kirk et al., 2024; Jiang et al., 2026), arguing that pre-training makes models broadly competent in similar ways, and reinforcement learning with human feedback (used to develop frontier models) pushes them toward similar preferred answers, with output variation driven to a considerable degree by prompting. In other words, there are strong theoretical reasons to expect model errors to be correlated—and they are in practice, as documented on the benchmark datasets used to evaluate frontier models (Kim et al., 2025).

In short, agreement across AI-generated measurements can simply reflect shared architectures and underlying training data, rather than convergence to implementing the construct as the researcher intends. At the same time, because the space of implementation choices is high-dimensional (the prompt space, for instance, is effectively infinite), there are also many levers that when varied can produce diverging responses for the same construct.

Where does this leave the researcher? The choice among models, prompts, and implementation details is consequential, and no particular choice has a privileged claim to the construct. The researcher can therefore search across models and prompts, deliberately or not, until the downstream estimate looks right: p-hacking in the form of measurement choice. Empirical economics has confronted a version of this problem before. When computation became cheap, researchers could search across regression specifications until a result emerged (e.g., Sala-I-Martin, 1997). The response was to build causal inference toolkits that guide

and discipline the researcher’s choice of specification. We are now building toolkits that provide the same discipline for AI-generated measurements.

3.3 Routes to Credible Measurement

Consider again the researcher who converts a rubric into a prompt and asks an LLM to label every document in the corpus. Section 3.2 ruled out two easy answers: treating the labels as the measurements themselves, and defending them simply with a few sensitivity checks across alternative AI-generated measures. Two routes remain. The first treats the model as a black box and establishes credibility by validating its outputs. The second opens the black box far enough to justify the assumptions about model behavior that identification requires.

There is a strong case for pursuing the first approach when it is feasible, because the errors made by deep neural networks can shift in potentially complex and meaningful ways with the distribution of inputs and implementation details. This can be difficult to model. Suppose our hypothetical researcher drew a random sample of documents from the corpus and carefully labeled each document according to their rubric. On this validation sample, the researcher observes both the measurement that aligns with the desired implementation of the rubric and the AI-generated measurement, and so can learn how the model errs relative to their intended construct and correct the downstream estimates accordingly.

This approach requires no assumptions about how the specific AI model produces its outputs, and that is a central virtue: the model has millions to billions of parameters that were learned in a complex way on data the researcher has never seen. The AI-generated measurements need not be unbiased, or even accurate (although inaccurate measurements will lead to wider confidence intervals). Instead, their errors are estimated from a random validation sample and corrected. This idea is an old one in the measurement error literature ([Bound and Krueger, 1991](#); [Lee and Sepanski, 1995](#); [Chen, Hong and Tamer, 2005](#); [Chen, Hong and Nekipelov, 2011](#)), and it has recently been revived for AI-generated measurements ([Angelopoulos et al., 2023](#); [Egami et al., 2024](#); [Carlson and Dell, 2025](#); [Ludwig, Mullainathan and Rambachan, 2026](#)).

Using a validation sample in this manner also clarifies the researcher’s purpose for using AI. The model no longer defines the measurement; it scales an existing one. The measurement process the researcher would defend (for example, a documented rubric with adjudicated edge cases applied by trained experts) cannot reasonably cover a massive corpus of documents. Within this framework, the existing measurement only needs to cover an affordable, high-quality validation sample, and the AI scales it across the rest of the corpus, augmenting researcher-implemented measurements at their best. When regressions became cheap, the

limiting factor became credible identification strategies ([Angrist and Pischke, 2010](#)). Now that measurement costs are falling, the limiting factor is increasingly having a high-quality measurement process in the first place. Section 4 provides an introduction to this framework.

This is not, however, the only route. In some settings, the researcher may not wish to privilege a validation sample, instead viewing all measurements as noisy attempts to capture an underlying latent concept. In others, collecting high-quality measurements in the population under analysis is prohibitively costly. Validation data may exist, but for a different study, region, or period. In both cases, identification rests on assumptions about the behavior of the AI models themselves, and those assumptions bring both new challenges and tools. Section 5 examines these settings.

Finally, there are cases where collecting random validation data is impossible, and the conditions required to achieve identification under assumptions about AI model behavior fail. There may still be strong reasons for using AI measurements to study a fundamentally important question and other avenues may exist for suggestive validation. However, it is necessary to proceed with an awareness of the potential pitfalls elaborated above.

4 Observation with Validation Samples: A Missing-Data Perspective

4.1 Inference with AI Predictions as Inference with Missing Structured Data

We now develop the validation route introduced above. We would like a broadly applicable framework that corrects the potential biases introduced by AI measurement, yielding estimates of target parameters that are robust to the choice of measurement instrument (*e.g.*, different models, different prompts, different implementation details). At the same time, better measurements should improve precision, thereby clarifying the cost-benefit tradeoff associated with investments in measurement quality.

Foundational work on semiparametric inference with missing data, and in particular the [Rubin \(1976\)](#) missing-at-random (MAR) mechanism, provides just such a framework. It may at first seem surprising to frame inference with AI predictions as a missing-data problem, but MAR is highly suitable because researchers often lack the low-dimensional summaries of unstructured data that are needed for statistical analysis. For instance, a researcher may have newspaper articles and use an LLM to impute whether these articles discuss economic policy uncertainty, the variable of interest. Because the articles do not include these binary

summaries, the problem is fundamentally one of missing data, which the researcher addresses by using an LLM as a scalable imputation technology.

Rubin’s seminal work is closely connected to large, subsequent literatures on semiparametric inference with missing data, measurement error, causal inference (a closely related missing data problem where counterfactual outcomes are missing), and inference with black-box AI predictions. Semiparametric missing-data frameworks are particularly complementary to deep neural networks because they allow the data to speak as much as possible, placing minimal assumptions on the deep neural network. This review follows the treatment in [Carlson and Dell \(2025\)](#), which applies Rubin’s MAR framework to a variety of settings that are common in inference with AI predictions in economics. Readers are referred to this article for elaboration on how their framework, referred to as MAR-S (for **Missing At Random Structured Data**), relates to the aforementioned literatures.

The core idea of missing-data frameworks is to use a validation sample to estimate the difference between imputed data and high-quality labels, adjusting target estimates accordingly. Validation data derive from an implementable, well-specified construct definition stated by the researcher. They are obtained through non-scalable, methodically applied processes, such as expert annotation or measurement by scientific ground instruments. The construct may be a highly flawed attempt to capture the concept of interest, but at the observation stage, it is taken as given.

The central assumption in the MAR-S framework—as the name underscores—is the missing at random assumption: after adjusting for observables, annotated and unannotated observations are comparable in their ground truth values. In other words, there are no unaccounted confounders determining whether an instance of unstructured data is in the validation sample.

4.2 The Formal Framework

MAR-S recasts robust and efficient inference with high-dimensional, unstructured data as inference on missing low-dimensional structured data. Structured data, $M \in \mathcal{M}$, are low-dimensional variables that can be used directly in estimating equations, whereas unstructured data, $U \in \mathcal{U}$, are high-dimensional and generally unsuitable for direct use in estimation.

Low-dimensional structured data are observed through “annotation.” Because this process is too expensive to scale to the full dataset, the researcher estimates an imputation function, $\hat{\mu}$, to impute the missing structured data. This allows the researcher to leverage the full unstructured dataset, which is often orders of magnitude larger than the validation sample. Increasingly, deep neural networks serve as this imputation function.

Rubin’s missing at random mechanism is closely linked to the Rubin Causal Model (Rubin, 1974; Imbens and Rubin, 2015), and accordingly we use potential outcomes notation. We observe a random variable $M \in \mathcal{M} \subseteq \mathbb{R}$ that is subject to missingness, with annotation status indicated by $A \in \{0, 1\}$. We write

$$M = AM^{a=1} + (1 - A)M^{a=0}.$$

We set $M^{a=0} = 0$ with probability one and, for notational convenience, define $M^* := M^{a=1}$. It follows that

$$M = AM^*.$$

We refer to M^* as the “ground truth” potential outcome and to A as the “annotation indicator,” which indicates whether an instance is in the validation sample. We use the term ground truth because it is overwhelmingly used in the literature to denote values in validation samples. To re-emphasize, this does not mean fundamental truth but simply the measurement that derives from careful implementation of a well-specified construct.

The first assumption is consistency of potential outcomes, which requires annotation status to be well defined and the label for any given instance to depend only on its own annotation status rather than on the annotation status of other instances. In practice, this means the researcher should apply the same measurement criteria, in the same way, to all data instances in the validation sample.

The central identifying assumption is missing at random. For ground truth potential outcome $M^* \in \mathcal{M}$, annotation indicator $A \in \{0, 1\}$, observed covariates $X \in \mathcal{X}$, and unstructured data $U \in \mathcal{U}$, we assume

$$(U, M^*) \perp\!\!\!\perp A \mid X.$$

After adjusting for observables X , annotated and unannotated data are comparable in their ground truth values. This assumption is analogous to “selection on observables” in causal inference, a closely related missing data problem where counterfactual outcomes are missing.

We define the “annotation score function” as

$$\pi(x) := P(A = 1 \mid X = x).$$

The baseline assumption is that $\pi(x)$ is fixed, known, and bounded away from zero. This embeds the “strong overlap” assumption commonly imposed in observational causal inference settings.

The annotation score will generally be known when the researcher designs the annotation

process—and should be documented when data are made publicly available for download by others—but this assumption can be relaxed when this is not the case.

The strong-overlap condition can also be relaxed to settings with decaying overlap. This scenario is relevant to massive data, where it is impossible for the researcher to validate more than a tiny share of a dataset that may contain millions or even billions of observations. As [Kallus and Mao \(2025\)](#) make clear in their analysis of treatment effects with surrogate outcomes—a problem that parallels AI inference in its underlying structure—there is no need to change the MAR-S estimator itself under an asymptotic regime in which the number of annotations diverges while the ratio of labeled to unlabeled observations converges to zero. The appropriate asymptotic efficiency analysis shows that the variance of the estimator is driven primarily by the size of the labeled dataset rather than by the size of the full dataset.

Additionally, efficiency requires mean squared error consistency:

$$E \left[\left(\hat{\mu}(\tilde{X}) - \mu(\tilde{X}) \right)^2 \right] = o(1).$$

Intuitively, the expected squared error of the imputation estimator should converge to zero as the amount of data used to train it grows. This condition is needed only for efficiency. It is *not* necessary for unbiased inference. It is relatively mild in the context of deep neural networks.⁴

[Carlson and Dell \(2025\)](#) apply this framework to derive estimators for a variety of common empirical scenarios in economics—e.g., ordinary least squares, instrumental variables, difference-in-differences—when left-hand side variables, right-hand side variables, or both are imputed with AI. A package is available for implementing these debiased estimators.⁵

The intuition for MAR-S can be illustrated by deriving the debiased estimator for a mean. Following [Chen, Hong and Tarozzi \(2008\)](#), who provide general results on semiparametrically efficient estimation of parameters identified by moment conditions with missing data, the moment function for the MAR-S debiased mean is simply $M^* - \theta$. The corresponding estimator is

$$\hat{\theta} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(\frac{A_i}{\pi(X_i)} \left[M_i - \hat{\mu}(\tilde{X}_i) \right] + \hat{\mu}(\tilde{X}_i) \right),$$

where $\mathcal{I}$ is the set of indices allocated to the estimation partition and $\hat{\mu}$ is an approximation to μ learned independently of the estimation sample, either because it is fixed in advance

⁴If the annotation score function must itself be estimated, the researcher must then assume that the imputation and annotation score functions converge at sufficiently fast rates ([Kennedy, 2023](#)), for example at $n^{-1/4}$ -type rates. While convergence rates are unknown given their massive-scale pre-training, these networks appear to converge quickly when being tuned to perform a particular well-defined task.

⁵<https://github.com/jscarlson/mar-s>

(*e.g.*, by an off-the-shelf model) or because it is fit on a separate tuning split. This is the augmented inverse probability weighted (AIPW) estimator, with M^* playing the role of a potential outcome.

To better understand the intuition, suppose that

$$\pi(X_i) = \frac{|\mathcal{I} \cap \mathcal{J}|}{|\mathcal{I}|},$$

where $\mathcal{J}$ is the set of indices corresponding to annotated observations (in other words, a given share of the instances are randomly annotated). Then the estimator can be written as

$$\hat{\theta} = \underbrace{\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \hat{\mu}(\tilde{X}_i)}_A + \underbrace{\frac{1}{|\mathcal{I} \cap \mathcal{J}|} \sum_{i \in \mathcal{I} \cap \mathcal{J}} (M_i^* - \hat{\mu}(\tilde{X}_i))}_B.$$

Term A is the imputation-based estimate of $E[M_i^*]$ in the estimation sample—what we would get if we ignored the potential measurement error in the neural network predictions—while term B is a bias-correction term that estimates the measurement error of the imputation function in the annotated sample. This expression is reproduced in recent work on “prediction-powered inference” ([Angelopoulos et al., 2023](#)). In a big data world, the variance of term A will be small, but the variance of term B could be quite large if AI predictions deviate substantially from M^* : noisier predictions will lead to noisier estimates of the target parameter.

Other estimators are largely analogous to the mean case, whether the imputed variable is an outcome, an instrument, a conditioning variable, etc., and interested readers are referred to [Carlson and Dell \(2025\)](#) for derivations. Two takeaways follow from the general treatment. First, estimators derived under MAR-S, and under related missing-data frameworks, are Neyman orthogonal: they are insensitive to first-order errors in the imputation function, so an inaccurate $\hat{\mu}$ does not compromise the validity of the estimate. Second, accuracy still matters, but for precision rather than validity. The variance of the estimator falls as $\hat{\mu}$ improves, so the imputation function should be learned as accurately as possible on a held-out sample. In other words, AI predictions, even arbitrarily bad ones, do not endanger a properly debiased estimate; the more accurate the predictor, the more precise the conclusions that can be drawn.

4.3 Aggregating AI Predictions

A key limitation of debiasing frameworks is that they assume that validation data are available for the imputed variables used in the estimating equation. In many applications in applied economics, however, validation data are available only at the granular level of individual texts or images, whereas the variable of interest is a potentially nonlinear function or functional of the granular missing data.

To make debiasing practically useful for many economic applications involving aggregated and transformed AI predictions, [Carlson and Dell \(2025\)](#) address this mismatch between the level at which ground truth is observed and the level at which the variable enters the estimating equation.

Fortunately, the solution is straightforward to implement in many scenarios. Consider the linear model

$$Y_i = X_i^* \beta + \varepsilon.$$

where X_i^* is an aggregate that is not actually observed, but is instead a function or functional of AI predictions. For example, X_i^* might be mean economic policy uncertainty in year i , derived from predictions at the level of individual news articles that are aggregated to the annual level.

Crucially, valid debiasing—while it does not remove measurement error—transforms any systematic biases from the AI predictions into classical measurement error. Hence, the approach is to first construct a debiased aggregate using MAR-S, and then proceed with the standard method-of-moments correction for classical errors-in-variables ([Deaton, 1985](#); [Fuller, 1987](#)).

This setup extends readily to a number of common empirical settings, including clustering and panel data ([Deaton, 1985](#)). It can also accommodate heterogeneity in the variance across aggregated observations, cases in which the outcome is itself generated by a MAR-S first step, and non-normal measurement error distributions. [Carlson and Dell \(2025\)](#) apply it to an analysis from [Baker, Bloom and Davis \(2016\)](#) where predictions of whether individual news articles discuss economic policy uncertainty are aggregated to the national annual level, logged, differenced, and interacted with a firm-level characteristic. Classical measurement error frameworks can accommodate these transformations. Importantly, using MAR-S to construct a debiased annual mean, rather than simply treating errors from the neural network as classical, is crucial for making the classical measurement error model plausible.

4.4 Annotation in Practice

In the most straightforward cases, researchers can annotate or access validation data for a simple random sample of data instances. However, in other cases, a simple random sample is unlikely to be informative. In large datasets, the structured data of interest may correspond to a rare event. For instance, a researcher may be interested in social media posts about inflation, which are a very small share of the total corpus of social media posts. In the machine learning literature, this is referred to as “class-imbalanced data.”

In such a setting, social scientists often use keyword-based filtering to select texts for annotation: texts containing specific keywords receive some positive probability of annotation, while all other texts are excluded. This approach violates strong overlap when structured data are then imputed from the full corpus, because it assigns zero annotation probability to some data instances. Intuitively, a language model’s measurement error is plausibly systematically correlated with the terms that appear in the text, and so annotating only articles with particular terms may not be informative about how the model performs on a broader variety of texts.

A researcher could limit attention to only texts with certain keywords, but in a variety of cases this risks overlooking many relevant texts, leading to a form of selection bias. Fortunately, there are alternatives to keywords for annotating class-imbalanced data. Existing work in machine learning and statistics, which can be incorporated into the MAR-S framework, suggests that researchers should oversample instances that are harder to impute (Zrnic and Candès, 2024). This aligns closely with a broad literature on importance sampling (Owen, 2013). But this prescription cannot be implemented directly because the difficulty of imputation depends on the unknown ground truth. However, the machine learning literature suggests various observable proxies for imputation difficulty, including model-based variance estimates, cross-validation residuals, and disagreement across ensemble predictions. Another approach that can work well in practice is to calculate the distance in embedding space to some relevant query—*e.g.*, this post is about inflation—where this embedding distance serves as the observable X that determines annotation probability in the MAR-S framework (see Dell, 2025, for an elaboration on embedding models and distances).

Another frequent question is how large the validation sample should be. The number of effective observations represented by the combined validation and imputed data depends on the asymptotic correlation between the validation sample-only estimator and the imputation-only estimator. For example, in the case of mean estimation, Broska, Howes and Van Loon

(2025) define the effective number of observations as

$$n_0 := |\mathcal{I} \cap \mathcal{J}| \times \frac{|\mathcal{I}|}{|\mathcal{I} \cap \mathcal{J}| + |\mathcal{I} \cap \mathcal{J}^c| (1 - \max\{\tilde{\rho}^2, 0\})},$$

where $\mathcal{J}$ is the set of annotated indices, $\mathcal{J}^c$ is the set of unannotated indices, and $\tilde{\rho}$ is the asymptotic correlation between the ground truth only estimator and the imputation-only estimator,

$$\tilde{\rho} := \text{Cor}(M^*, \mu(\tilde{X})).$$

The more predictive the imputation function, the greater the effective sample size. In the limiting case of a perfect imputation function, the effective sample size equals the full size of the dataset.

The asymptotic correlation between the ground truth-only estimator and the imputation-only estimator is unknown, so this cannot be calculated directly. Researchers may nevertheless have a prior about it, or wish to calculate effective sample sizes under different scenarios. For a binary classifier (e.g., “is this post about inflation?”), in practice often a couple hundred high-quality annotations are sufficient when the imputation function is reasonably accurate. As predictions become noisier, incorporating them contributes less and less relative to simply using the small validation sample. In this scenario, it is necessary for the researcher to invest in more accurate AI predictions (by training a custom model, using higher quality training data or a larger model, etc.) if they wish to leverage AI predictions to improve power.

What about errors in the validation sample? One cannot correct arbitrary measurement error in AI predictions using validation data measured with arbitrary error, although it is possible to make some progress by putting structure on the errors. To minimize errors, validation data should be the researcher’s measurement at its best: a modestly sized sample built using a carefully defined construct by a process that is tightly supervised by the researcher. AI can scale the measure, so large samples are not needed. Rather, the job of the researcher is to ensure the validation data are of the highest quality. If collecting high-quality validation data is not feasible, a different approach to measurement with AI predictions is required.

5 Observation without Random Validation Samples

In Section 4, we saw that the missing at random assumption is analogous to the selection-on-observables assumption in causal inference, which can likewise be viewed as a missing data problem: treatment assignment reveals which potential outcome is observed, whereas

in our setting annotation determines whether M^* is observed. Hence, much as randomized treatment is the gold standard for causal inference, random validation samples are the gold standard for inference with unstructured data. Yet, it is often impossible to create a high-quality, random validation sample, much as randomized treatment is often infeasible in causal inference. Fortunately, just as causal inference can proceed under additional assumptions when randomized treatment assignment is infeasible, valid inference with unstructured data can proceed under additional assumptions when a random validation sample cannot be constructed.

The validation sample framework in Section 4 rests on two requirements. First, the researcher must specify a measurement process that they are prepared to defend: an explicit rubric or instrument-based measurement that defines the construct M^* and adjudicates edge cases. Second, the researcher must apply that measurement process to a random validation sample drawn from the population under analysis. This section considers scenarios in which each requirement fails.

Sometimes there are no existing validation data, and the researcher cannot create them if no technology can access M^* —that is, all measurement technologies are inherently noisy—or applying a technology that can is infeasible (Section 5.1). Other times validation data exist or can be compiled, but from another region, another period, or another population, so the labels are not a random validation sample for the analysis at hand (Section 5.2). In both cases, identification of the downstream parameter requires additional measurement assumptions, and those assumptions are claims about the behavior of frontier AI and machine learning models.

Why might we be in a setting with non-random or non-existent validation data? A useful distinction that often matters in practice is “ U -measurability.” M^* is U -measurable if it can be accessed solely from unstructured data U without consulting any external information. For example, simple text analysis tasks are often U -measurable. If a construct is U -measurable, it will typically be feasible for the researcher to construct a random validation set if they can access U (which may not be the case for externally compiled data). There are a wide variety of settings, however, where M^* is not U -measurable. This is particularly common in remote sensing. An annotator cannot directly observe crop type from lower resolution satellite data, or consumption from nighttime lights. Instead, validation requires ground measurements on crop type or consumption that are very costly to collect. Often some validation sample exists—without labeled data on crop type or consumption, a model could not have been trained to predict them in the first place. But typically the researcher wants to study a different region, time period, or population. Sometimes, the entire motivation for using remote sensing is to collect data for areas where obtaining ground mea-

surements is impossible (*e.g.*, remote or conflict prone areas). In short, creating a random validation sample for the population of interest is often out-of-reach.

5.1 Multiple Measurements

Suppose we are in a world where instead of some non-scalable technology revealing M^* , the best each measurement technology j can do is to access $M^* + \eta_j$, where η_j is an error term associated with measurement technology j . The construct is still well-defined, but it can only be applied with noise. (If the construct definition itself is uncertain, a discovery regime, as described in Section 2.1, will plausibly be more suitable.)

In this case, the researcher may instead wish to treat the construct M^* as if it were latent and each measurement protocol were a repeated, noisy measurement of it. Concretely, suppose the researcher forms J measurement protocols, producing measurements $\widehat{M}_1, \dots, \widehat{M}_J$ of the same construct. Latent variables observed only through error-prone measurements are the subject of a rich literature on measurement systems and finite mixtures, stretching from classical models of observer error to modern nonclassical measurement error (Dawid and Skene, 1979; Schennach, 2020, 2021).

Might these identification results be useful here? This depends on whether additional assumptions about the AI-generated measurements are met.

Consider the simple case in which the construct is binary, $M^* \in \{0, 1\}$, and the researcher generates at least three measurements, $J \geq 3$. Classical results dating to Kruskal (1977) and Dawid and Skene (1979) show that if the measurement errors are conditionally independent across protocols given the intended construct, and each protocol j is informative (meaning its true positive rate exceeds its false positive rate: $\Pr(\widehat{M}_j = 1 \mid M^* = 1) > \Pr(\widehat{M}_j = 1 \mid M^* = 0)$), then the joint distribution of $(\widehat{M}_1, \dots, \widehat{M}_J)$ identifies both the distribution of the latent construct M^* and each protocol’s error rates. With enough conditionally independent measurements of the same latent construct, the data themselves reveal how often each protocol errs. No validation sample is required.⁶

When such identification results are applied to AI, the assumptions are not opaque conditions, but rather statements about the behavior of AI models. The informativeness condition is plausibly mild. As long as the researcher is measuring something sufficiently concrete and well-defined, it seems reasonable to assume the AI predictions are better than chance.

The problem is justifying that measurements are conditionally independent across models, prompts, and other implementation details. Conditional independence asserts that different models or different prompts err independently on the same input. There is every

⁶See Hu and Schennach (2008), Allman, Matias and Rhodes (2009), and Bonhomme, Jochmans and Robin (2016) for generalizations of these ideas to richer constructs, measurements, and error structures.

structural reason to expect the opposite: frontier models share architectures, overlap heavily in their training corpora, and are tuned with similar human-feedback pipelines (Kleinberg and Raghavan, 2021; Bommasani et al., 2022). In some cases, they are directly “distilled” from—or trained on data derived from—other frontier models.

Because conditional independence is a claim about how frontier models err, we can empirically check whether it is true if we are able to access M^* . As argued in Section 3, a natural place to look is the diverse set of benchmark datasets that are used to evaluate frontier models. Researchers have documented that frontier models err on the same items, and more capable models err more similarly (Kim et al., 2025). The same behavior appears in economic measurement, where errors are strongly correlated across models and prompts (Ludwig, Mullainathan and Rambachan, 2026; Chen, Rambachan and Tamer, 2026).

The failure of conditional independence does not mean that multiple measurements are uninformative. It means the assumption must be relaxed. One path allows for bounded violations of conditional independence, with the size of the permitted violation calibrated to external evidence on how correlated model errors actually are in some external dataset(s) for which M^* is available (*e.g.*, AI benchmarks or economic domains where validation data exist). The researcher then reports an identified set for the downstream parameter rather than a point estimate, together with a breakdown value: how large would the violation have to be to overturn the conclusion? Chen, Rambachan and Tamer (2026) develop this approach for AI-generated measurements, and we refer the interested reader there for further details. The general lesson is that, given what we know about how these models err, multiple measurements may partially identify downstream parameters, with bounds whose width is governed by that behavior.

To the extent that all measures are generated by a black-box AI model, we might also worry about the construct validity issues raised in Section 2: how do we know the models are giving multiple measurements of, *e.g.*, trust, rather than something that is simply correlated with it? Here, evidence on convergent and discriminant validity remains useful.

The measurement error literature offers many identification arguments beyond repeated measurements: for example, based on instrumental variables, auxiliary data, and restrictions on the joint distribution of errors (see Chen, Hong and Nekipelov (2011); Schennach (2020, 2021) for reviews). Historically, the assumptions behind these arguments concerned human measurements. Applied to AI-generated measurements, the same assumptions become properties of models, and models can be queried, tested, and benchmarked. Assessing which identification arguments are credible, and how sensitive conclusions are to violations of their assumptions, becomes an empirical exercise: ask what each assumption asserts about the behavior of frontier AI models, and bring evidence to bear on it.

5.2 Data Combination

We next turn to the second failure of the validation sample framework, in which annotated data exist but not for the population under analysis. Remote sensing is the showcase application. Machine learning models applied to satellite imagery and other digital traces now produce measurements of poverty, living standards, deforestation, weather and climate patterns, and crop cover across the globe, including places that may be difficult for survey teams to reach or where ground measurement instruments are difficult to maintain (Hansen et al., 2013; Blumenstock, Cadamuro and On, 2015; Jean et al., 2016; Rolf et al., 2021; Aiken et al., 2025). These measurements increasingly serve as outcomes and regressors in development, environmental, and urban economics. In these applications, ground truth comes from the traditional measurement processes that the machine learning model seeks to scale: an in-person survey that elicits consumption or income, surveyors sent into forests to measure tree cover, or physical monitors that record pollution. When a conditionally random validation sample drawn from the population under analysis exists, the framework of Section 4 applies directly (Proctor, Carleton and Sum, 2023; Kluger et al., 2025; Lu et al., 2025; Pelletier et al., 2026).

Because collecting such data is costly, the machine learning model is commonly trained and validated on an auxiliary sample where ground truth exists, in one country, one period, or one sensor’s footprint, and then deployed on the analysis sample, where ground truth has not been collected. The auxiliary sample is not a validation sample. Nothing guarantees that the relationship between predictions and ground truth transfers, and predictive performance often degrades when transferring across regions and periods (e.g., Proctor et al., 2025). More generally, a central limitation of neural networks is that performance tends to decline with domain shift away from the training data distribution (Ben-David et al., 2010).

Often, validation data do not exist precisely because regions are fundamentally different (e.g., more geographically isolated, poorer, or conflict prone) from areas where ground data collection is feasible. These concerns are compounded when, as is often the case, the only available evaluation sample is a held-out split from the same annotated dataset used for model training, since performance on a substantively different target population may shift in complex ways relative to performance on the in-sample distribution.

To proceed, the researcher must assume that some feature of the auxiliary sample carries over to the analysis sample. If this is not plausible, researchers can consider whether the multiple measurement framework is a better fit.

The literature has organized around two alternative identifying assumptions for this data combination problem. Let a denote the auxiliary sample, in which both the AI-generated measurement $\widehat{M}$ and the ground truth M^* are observed; let e denote the analysis sample, in

which only $\widehat{M}$ is observed; and let P_a and P_e denote the corresponding distributions.

The first identifying assumption, which we call *outcome stability*, states that the distribution of the ground truth given the prediction is the same in both samples: $P_a(M^* | \widehat{M}) = P_e(M^* | \widehat{M})$. In the auxiliary sample, the researcher learns the conditional distribution of the ground truth measurement given the prediction, and this conditional distribution transfers to the analysis sample. Outcome stability may be plausible when the prediction was constructed to forecast the outcome, as when a poverty score is built from phone metadata. It is the same assumption underlying the literature on “surrogate outcomes” (Chen, Hong and Tarozzi, 2008; Kallus and Mao, 2025; Athey et al., 2026).

To see what outcome stability delivers, consider estimating the prevalence of a binary construct in the analysis sample, $\theta = P_e(M^* = 1)$. Under outcome stability, the researcher can take the conditional distribution learned in the auxiliary sample, apply it to each prediction in the analysis sample, and average: $\theta = \mathbb{E}_e[P_a(M^* = 1 | \widehat{M})]$. The argument parallels the validation sample analysis of Section 4, in which the researcher also learns the relationship between the ground truth and the AI-generated measurements on an annotated subsample and extrapolates to the rest of the sample. The difference is that random annotation guarantees the extrapolation is valid by design, whereas here it must be assumed. Making this argument precise requires additional conditions; we refer the reader to existing work for formal statements (Kallus and Mao, 2025; Athey et al., 2026).

The second identifying assumption, which we call *measurement stability*, states that the distribution of the prediction given the ground truth is the same in both samples: $P_a(\widehat{M} | M^*) = P_e(\widehat{M} | M^*)$. In the auxiliary sample, the researcher learns how the AI-generated measurement errs given the truth, that is, the model’s error rates, and these error rates transfer to the analysis sample. Measurement stability may be plausible when the outcome physically generates the signal that the model reads, as when crop burning produces the smoke plume that a satellite sensor detects. Rambachan, Singh and Viviano (2026) study identification under measurement stability for AI-generated measurements in remote sensing.

Under measurement stability, the construct’s prevalence is identified by a different argument (Rambachan, Singh and Viviano, 2026). The rate of positive predictions in the analysis sample mixes together the model’s error rates: $P_e(\widehat{M} = 1) = \theta P_a(\widehat{M} = 1 | M^* = 1) + (1 - \theta) P_a(\widehat{M} = 1 | M^* = 0)$. Provided the predictions are informative, in the same sense as in Section 5.1, the researcher can invert this equation, correcting the analysis sample’s prediction rate using the auxiliary sample’s error rates. Notice that the two stability assumptions use the same data but produce different formulas, and in general they produce different answers.

Which stability assumption should the researcher invoke, if either? The substance of the

choice is which conditional distribution is plausibly invariant across the two samples. Measurement stability asserts that the way the measurement is generated from the underlying construct is stable across the samples, for example, the way crop burning or deforestation appears in satellite images. It therefore faces engineering threats, such as changes in sensors, image resolution, or model versions between the samples. Outcome stability asserts that the relationship between the construct and the prediction is stable, and so anything that alters the distribution of the construct given the prediction across the samples is a potential threat. Stating which threats are plausible in the application at hand is the analogue, in this setting, of defending an identifying assumption in a research design. The choice is worth making carefully: exploiting the correct stability assumption can deliver substantially more precise estimates than discarding the auxiliary sample or applying the predictions naively (Rambachan, Singh and Viviano, 2026; Alsharif et al., 2026).

Like conditional independence in Section 5.1, both measurement stability and outcome stability are assumptions about the behavior of frontier AI and machine learning models, in this case about how that behavior travels across settings. These too are empirical claims, and benchmarks are again the natural place to assess them. The existing benchmarks, however, measure a different property: they document how predictive performance degrades across regions, periods, and sensors. But predictive performance does not reveal whether outcome stability or measurement stability holds. Building benchmarks that track whether these conditional distributions transfer, on the measurement tasks economists care about, is a natural next step and would put the choice between stability assumptions on an empirical footing.

Measurement without a random validation sample is not measurement without discipline. The discipline instead moves from the design of the validation sample to the defense of the assumption. Building that discipline, through new identification arguments and new evidence on how models err, is an active area of research. The researcher’s obligations follow: state the assumption explicitly, defend it in the application at hand, evaluate it where evidence exists, and report how conclusions change as it is relaxed.

6 Conclusion

Empirical economics is limited by what we can measure. Structured data record only features someone chose to observe, while careful, interpretable measurement is often too costly to implement at scale. AI relaxes both constraints. Unstructured data capture dimensions of economic life that have rarely or never been measured systematically, and AI can transform them into structured variables, allowing measurement processes once feasible only on

small samples to be applied across entire corpora. Opportunities to apply AI arise throughout the measurement pipeline, from discovering candidate constructs to defining them and implementing them at scale. AI is therefore widening the aperture of what economics can measure.

This article focuses primarily on observation, where AI usage is by far the most widespread and where the associated statistical challenges are most developed. The central idea is simple: apply a defensible measurement process to a random validation sample and compare these measurements to the AI generated measures to correct downstream estimates. This logic predates AI: statistical agencies, for example, have long used expensive, high-quality validation samples to correct much cheaper census measurements collected across the full population.

This framework is particularly well suited to black-box AI because researchers need not model the neural network, assume its predictions are unbiased, or even require them to be accurate. Instead, prediction errors are learned from a validation design the researcher controls and should therefore be able to interpret and defend. Even a model with billions of parameters trained on inaccessible data can thus support credible, interpretable empirical estimates. When random ground truth cannot be collected, however, stronger assumptions become unavoidable. This literature is still developing, and an important direction for future work is to connect evaluations of frontier models more directly to the assumptions required for credible empirical research.

In Appendix A, we provide a checklist for researchers using AI-generated variables. It is designed both to help authors apply the framework developed in this article and to guide transparent reporting. It identifies which of the three cases a study falls into, clarifies the relevant assumptions and requirements, and specifies what should be reported about the measurement process—including construct definition, validation-label generation, and the model and (if applicable) prompt used. We hope it promotes more rigorous and transparent empirical use of AI-generated measurements.

References

- Aiken, Emily, Suzanne Bellue, Joshua E. Blumenstock, Dean Karlan, and Christopher Udry. 2025. “Estimating Impact with Surveys versus Digital Traces: Evidence from Randomized Cash Transfers in Togo.” *Journal of Development Economics*, 175: 103477.
- Allman, Elizabeth S., Catherine Matias, and John A. Rhodes. 2009. “Identifiability of Parameters in Latent Structure Models with Many Observed Variables.” *The Annals of Statistics*, 37(6A): 3099–3132.
- Alsharif, Haya, Ashesh Rambachan, Rahul Singh, and Davide Viviano. 2026. “Causal Inference with Satellite Imagery: A Comparison of Methods for Forest Conservation Data.” *AEA Papers and Proceedings*, 116: 87–91.
- Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnica. 2023. “Prediction-Powered Inference.” *Science*, 382(6671): 669–674.
- Angrist, Joshua D. 2022. “Empirical Strategies in Economics: Illuminating the Path From Cause to Effect.” *Econometrica*, 90(6): 2509–2539.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. “The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics.” *Journal of Economic Perspectives*, 24(2): 3–30.
- Asirvatham, Hemanth, Elliott Moksiki, and Andrei Shleifer. 2026. “GPT as a Measurement Tool.” National Bureau of Economic Research NBER Working Paper 34834.
- Athey, Susan, Raj Chetty, Guido W. Imbens, and Hyunseung Kang. 2026. “The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely.” *The Review of Economic Studies*, 93(4): 2284–2312.
- Atreja, Shubham, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2025. “What’s in a Prompt? A Large-Scale Experiment to Assess the Impact of Prompt Design on the Compliance and Accuracy of LLM-Generated Text Annotations.” *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media*, 122–145.
- Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. “Measuring economic policy uncertainty.” *The Quarterly Journal of Economics*, 131(4): 1593–1636.

- Baron, Jason, Will Dobbie, Richard Lombardo, Ashesh Rambachan, and Joseph Ryan.** 2026. “Human Decisions and Machine Predictions in Multistage Systems.” Working paper, June 29, 2026.
- Barrie, Christopher, Alexis Palmer, and Arthur Spirling.** 2024. “Replication for Language Models Problems, Principles, and Best Practice for Political Science.” https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf.
- Batista, Rafael M., and James Ross.** 2024. “Words that Work: Using Large Language Models to Generate and Refine Hypotheses from Text.” Working paper.
- Baumann, Joachim, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B Gruber, and Dirk Hovy.** 2025. “Large language model hacking: Quantifying the hidden risks of using llms for text annotation.” *arXiv preprint arXiv:2509.08825*.
- Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan.** 2010. “A theory of learning from different domains.” *Machine Learning*, 79: 151–175.
- Blumenstock, Joshua E., Gabriel Cadamuro, and Robert On.** 2015. “Predicting Poverty and Wealth from Mobile Phone Metadata.” *Science*, 350(6264): 1073–1076.
- Bommasani, Rishi, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang.** 2022. “Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?” Vol. 35.
- Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin.** 2016. “Non-parametric Estimation of Finite Mixtures from Repeated Measurements.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(1): 211–229.
- Borges, Jorge Luis.** 1998. “On Exactitude in Science.” In *Collected Fictions*. 325. New York: Viking.
- Bound, John, and Alan B. Krueger.** 1991. “The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?” *Journal of Labor Economics*, 9(1): 1–24.
- Broska, David, Michael Howes, and Austin Van Loon.** 2025. “The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations.” SSRN Working Paper 5133034.

- Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky.** 2022. “Computational Analysis of 140 Years of US Political Speeches Reveals More Positive but Increasingly Polarized Framing of Immigration.” *Proceedings of the National Academy of Sciences*, 119(31): e2120510119.
- Carlson, Jacob.** 2026. “Making Interpretable Discoveries from Unstructured Data: A High-Dimensional Multiple Hypothesis Testing Approach.” *arXiv preprint arXiv:2511.01680*.
- Carlson, Jacob, and Melissa Dell.** 2025. “A unifying framework for robust and efficient inference with unstructured data.” *arXiv preprint arXiv:2505.00282*.
- Chen, Xiaohong, Ashesh Rambachan, and Elie Tamer.** 2026. “Partial Identification from LLM Prompts.” June 25, 2026 version.
- Chen, Xiaohong, Han Hong, and Alessandro Tarozzi.** 2008. “Semiparametric Efficiency in GMM Models with Auxiliary Data.” *The Annals of Statistics*, 36(2): 808–843.
- Chen, Xiaohong, Han Hong, and Denis Nekipelov.** 2011. “Nonlinear Models of Measurement Errors.” *Journal of Economic Literature*, 49(4): 901–937.
- Chen, Xiaohong, Han Hong, and Elie Tamer.** 2005. “Measurement Error Models with Auxiliary Data.” *The Review of Economic Studies*, 72(2): 343–366.
- Chopra, Felix, and Ingar Haaland.** 2026. “Conducting Qualitative Interviews with AI.” Working paper, March 9, 2026.
- Coqueret, Guillaume, Joan Llull, Florian Oswald, Christophe Pérignon, Christoph Scheuch, and Lars Vilhuber.** 2026. “Randomness in large language models: What researchers need to know (and report).”
- Cronbach, Lee J, and Paul E Meehl.** 1955. “Construct validity in psychological tests.” *Psychological Bulletin*, 52(4): 281.
- Dawid, A. P., and A. M. Skene.** 1979. “Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm.” *Journal of the Royal Statistical Society Series C: Applied Statistics*, 28(1): 20–28.
- Deaton, Angus.** 1985. “Panel data from time series of cross-sections.” *Journal of Econometrics*, 30(1-2): 109–126.
- Dell, Melissa.** 2025. “Deep Learning for Economists.” *Journal of Economic Literature*, 63(1): 5–58.

- Donahue, Kate, Alexander Idarraga, Rohan Alur, Gabriel Brat, and Manish Raghavan.** 2026. “Leveraging signals in expert trace data for interpretable complementarity.”
- Donoho, David.** 2024. “Data Science at the Singularity.” *Harvard Data Science Review*, 6(1).
- Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei.** 2024. “Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses.” Working paper, November 17, 2024.
- Fuller, Wayne A.** 1987. *Measurement Error Models*. New York:Wiley.
- Geiecke, Friedrich, and Xavier Jaravel.** 2026. “Conversations at Scale: Robust AI-led Interviews.” Working paper, February 7, 2026.
- Haaland, Ingar K., Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart.** 2024. “Understanding Economic Behavior Using Open-ended Survey Data.” National Bureau of Economic Research NBER Working Paper 32421.
- Hansen, Matthew C., Peter V. Potapov, Rebecca Moore, Matt Hancher, Svetlana A. Turubanova, Alexandra Tyukavina, David Thau, Stephen V. Stehman, Scott J. Goetz, Thomas R. Loveland, Anil Kommareddy, Alexey Egorov, Louise Chini, Christopher O. Justice, and John R. G. Townshend.** 2013. “High-Resolution Global Maps of 21st-Century Forest Cover Change.” *Science*, 342(6160): 850–853.
- Hu, Yingyao, and Susanne M. Schennach.** 2008. “Instrumental Variable Treatment of Nonclassical Measurement Error Models.” *Econometrica*, 76(1): 195–216.
- Imbens, Guido W, and Donald B Rubin.** 2015. “Causal inference for Statistics, Social, and Biomedical Sciences.” *New York*, 517.
- Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Alampay Davis, David B. Lobell, and Stefano Ermon.** 2016. “Combining Satellite Imagery and Machine Learning to Predict Poverty.” *Science*, 353(6301): 790–794.
- Jiang, Liwei, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi.** 2026. “Artificial hivemind:

- The open-ended homogeneity of language models (and beyond).” *Advances in Neural Information Processing Systems*, 38.
- Kallus, Nathan, and Xiaojie Mao.** 2025. “On the Role of Surrogates in the Efficient Estimation of Treatment Effects with Limited Outcome Data.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2): 480–509.
- Kennedy, Edward H.** 2023. “Semiparametric doubly robust targeted double machine learning: a review.” arXiv:2203.06469.
- Kim, Elliot Myunghoon, Avi Garg, Kenny Peng, and Nikhil Garg.** 2025. “Correlated Errors in Large Language Models.” Vol. 267 of *Proceedings of Machine Learning Research*, 30038–30066. PMLR.
- Kirk, Robert, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu.** 2024. “Understanding the effects of rlhf on llm generalisation and diversity.” Vol. 2024, 20620–20653.
- Kleinberg, Jon, and Manish Raghavan.** 2021. “Algorithmic Monoculture and Social Welfare.” *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.
- Kluger, Dan M., Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates.** 2025. “Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling.”
- Kruskal, Joseph B.** 1977. “Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics.” *Linear Algebra and its Applications*, 18(2): 95–138.
- Lee, Lung-fei, and Jungsywan H. Sepanski.** 1995. “Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data.” *Journal of the American Statistical Association*, 90(429): 130–140.
- Liu, Haokun, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan.** 2025. “HypoBench: Towards Systematic and Principled Benchmarking for Hypothesis Generation.”
- Ludwig, Jens, and Sendhil Mullainathan.** 2024. “Machine Learning as a Tool for Hypothesis Generation.” *The Quarterly Journal of Economics*, 139(2): 751–827.
- Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan.** 2026. “Large Language Models: An Applied Econometric Framework.” *Annual Review of Economics*.

- Lu, Kerri, Dan M. Kluger, Stephen Bates, and Sherrie Wang.** 2025. “Regression Coefficient Estimation from Remote Sensing Maps.” *Remote Sensing of Environment*, 330: 114949.
- Modarressi, Iman, Jann Spiess, and Amar Venugopal.** 2025. “Causal Inference on Outcomes Learned from Text.”
- Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson.** 2025. “Sparse Autoencoders for Hypothesis Generation.”
- Movva, Rajiv, Smitha Milli, Sewon Min, and Emma Pierson.** 2026. “What’s in My Human Feedback? Learning Interpretable Descriptions of Preference Data.”
- Mullainathan, Sendhil, and Ashesh Rambachan.** 2025. “Science in the Age of Algorithms.” Working paper, October 10, 2025.
- Mullainathan, Sendhil, and Jann Spiess.** 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives*, 31(2): 87–106.
- Obermeyer, Ziad, Alexander Schubert, James Ross, Sendhil Mullainathan, and Markus Lingman.** 2026. “An ECG Biomarker for Sudden Cardiac Death Discovered with Deep Learning.” *Nature*. Published online 24 June 2026.
- Owen, Art B.** 2013. *Monte Carlo Theory, Methods and Examples*. <https://artowen.su.domains/mc/>.
- Pelletier, Johanne, Mira Korb, Solomon Alemu, Manex B. Yonis, Travis J. Lybbert, and Matthieu Stigler.** 2026. “Causal Inference with Predicted Outcomes: Correcting prediction error bias in satellite-based impact evaluation.” *Journal of Development Economics*, 179: 103655.
- Peng, Kenny, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg.** 2026. “Position: Use Sparse Autoencoders to Discover Unknowns.”
- Proctor, Jonathan, Tamma Carleton, and Sandy Sum.** 2023. “Parameter Recovery Using Remotely Sensed Variables.” National Bureau of Economic Research NBER Working Paper 30861.
- Proctor, Jonathan, Tamma Carleton, Trinetta Chong, Taryn Fransen, Simon Greenhill, Jessica Katz, Hikari Murayama, Luke Sherman, Jeanette Tseng, Hannah Druckenmiller, and Solomon Hsiang.** 2025. “What Can Satellite Imagery

and Machine Learning Measure?” National Bureau of Economic Research NBER Working Paper 34315.

Rambachan, Ashesh, Rahul Singh, and Davide Viviano. 2026. “Program Evaluation with Remotely Sensed Outcomes.”

Rolf, Esther, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. 2021. “A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery.” *Nature Communications*, 12: 4392.

Rubin, Donald B. 1974. “Estimating causal effects of treatments in randomized and non-randomized studies.” *Journal of Educational Psychology*, 66(5): 688.

Rubin, Donald B. 1976. “Inference and missing data.” *Biometrika*, 63(3): 581–592.

Sala-I-Martin, Xavier X. 1997. “I Just Ran Two Million Regressions.” *The American Economic Review*, 178–183.

Schennach, Susanne M. 2020. “Mismeasured and Unobserved Variables.” In *Handbook of Econometrics*. Vol. 7A, , ed. Steven N. Durlauf, Lars Peter Hansen, James J. Heckman and Rosa L. Matzkin, 487–565. Elsevier.

Schennach, Susanne M. 2021. “Measurement Systems.” Centre for Microdata Methods and Practice cemmap Working Paper CWP12/21.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. “Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting.”

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan. 2024. “Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function.”

Wang, Jenny S., Aliya Saperstein, and Emma Pierson. 2026. “In Your Own Words: Computationally Identifying Interpretable Themes in Free-Text Survey Data.”

Workman, Ben. 2025. “Inside the Black Box: Using Machine Learning to Discover Characteristics of Effective Teaching.” Working paper.

- Xiong, Chenfei, Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Lorena Calvo-Bartolomé, Alexander Hoyle, Zhijing Jin, Mrinmaya Sachan, Markus Leipold, Dirk Hovy, Mennatallah El-Assady, and Elliott Ash.** 2025. “Co-DETECT: Collaborative Discovery of Edge Cases in Text Classification.” 354–364. Association for Computational Linguistics.
- Yin, Michelle, Hoa Vu, and Claudia Persico.** 2026. “How (un)Stable Are LLM Occupational Exposure Scores? Evidence from Multi-Model Replication.” National Bureau of Economic Research NBER Working Paper 35110.
- Zhong, Ruiqi, Charlie Snell, Dan Klein, and Jacob Steinhardt.** 2022. “Describing Differences between Text Distributions with Natural Language.”
- Zhong, Ruiqi, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt.** 2023. “Goal Driven Discovery of Distributional Differences via Language Descriptions.”
- Zhou, Yangqiaoyu, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan.** 2024. “Hypothesis Generation with Large Language Models.”
- Zhu, Jian-Qiao, Hanbo Xie, Dilip Arumugam, Robert C. Wilson, and Thomas L. Griffiths.** 2026. “Using Reinforcement Learning to Train Large Language Models to Explain Human Decisions.”
- Zrnic, Tijana, and Emmanuel J. Candès.** 2024. “Active Statistical Inference.” *Proceedings of the 41st International Conference on Machine Learning*, 62993–63010.

A A Checklist for Observation with AI-Generated Variables

Purpose. This checklist helps authors report the measurement details a reader needs in order to assess AI-generated measures. We recommend either answering these questions explicitly in supplemental materials or using the checklist to confirm that its details are discussed somewhere in the article or replication materials. Answer to the best of your knowledge, and state “*unknown*” where a detail is not available, which is common for variables obtained from an external source. If a question is irrelevant, provide justification.

Provenance and Reproducibility

- ☐ Did you construct the AI-generated variable(s), or did you obtain them from an external source? If external, what is the source?
- ☐ If you constructed the AI-generated variable(s), will the code, prompts, and model versions be released no later than the time of publication, along with the data where permissible?
- ☐ Do the replication files include the outputs returned by the model, so that the measurements can be reproduced without re-querying it?

Construct Definition

- ☐ What is the definition of the construct(s) you measured with AI?
- ☐ Was any construct validation performed? If so, what evidence was used (*e.g.*, variables the construct is shown to correlate with, or to remain distinct from)?

Validation Regime

Identify which of the three regimes you are in. Complete only the matching sub-section below.

- ☐ Which validation regime applies: **(a) no validation data**, **(b) representative validation data**, or **(c) validation data available, but not for the sample of interest**?

(a) No validation data

- ☐ Why was obtaining validation data infeasible?
- ☐ What other evidence supports the validity of the AI-generated measure in the absence of a validation sample?

(b) Representative validation data

The validation labels are a random (conditional on observables) subsample of the observations used to generate predictions.

- ☐ How was the validation sample selected (*e.g.*, simple random sample, or at random conditional on covariates)? If covariates were used, what are they?
- ☐ Is the probability that an observation is annotated known by design, or was it estimated (and if estimated, how)? Did every type of observation have a positive probability of being annotated?

(c) Validation data available, but not for the sample of interest

*Ground truth exists, but it was collected in another sample (*e.g.*, another region, period, or population).*

- ☐ How do the two samples differ (*e.g.*, in the distribution of predictions, covariates, geography, or time period)? Which forms of shift matter most in your application?
- ☐ Which stability assumption do you rely on—such as outcome stability or measurement stability—and what evidence supports it (*e.g.*, comparisons of error rates, calibration, or predictive performance across the two samples)? If neither is defensible in your setting, say so.

Validation Dataset Description (if applicable)

- ☐ How large is the validation sample, and when were the labels collected?
- ☐ What technology produced the validation labels (*e.g.*, human annotators, instrument measurement)? Provide enough detail for a reader to understand how the measurement was implemented: for annotators, what training and instructions they were given; for instruments, the relevant technical specifications.
- ☐ Was a consistent rubric used to produce the validation labels?
- ☐ Was inter-annotator (or inter-instrument) agreement examined? If so, on how many observations, what was the agreement, and how were disagreements resolved?

Construction of the AI Predictions

- ☐ What model(s) were used, and were they off-the-shelf or fine-tuned? Give model versions and citations, where the models can be obtained, and for API-served or hosted models the access date and version string.
- ☐ How were the model and the prompt selected? What alternatives were considered, and were the reported estimates examined for sensitivity to those choices?

- ☐ If a prompt was used, what was it? Provide the full text, along with the generation settings for a language model (*e.g.*, temperature, top- p , maximum tokens, random seed, and number of samples drawn per observation).
- ☐ If training data were used to produce the predictions, how were the training data created? How large was the training sample? What were the relevant training details?
- ☐ How many observations have AI-generated measurements?
- ☐ Are the AI-generated measurements aggregated or otherwise transformed before use? If so, how, and at what level relative to the unit of analysis?

Use in Estimation

This checklist does not prescribe an estimator. The appropriate correction depends on the estimand, on the role the measured variable plays, and on the validation regime above.

- ☐ Does the AI-generated variable enter the analysis as an outcome, a regressor, an instrument, or in some other role?
- ☐ Do the reported estimates account for error in the AI-generated measurements? If so, what approach was used and what does it assume? If not, why is no correction needed for the estimand of interest?