A Measurement Science Roadmap: From Human Assessment to AI Evaluation

Sang Truong[1,2], Noah Goodman[1,3], Emma Brunskill[1], Ben Domingue[2],

Nick Haber[1,2*], Sanmi Koyejo[1*]

[1]Stanford Computer Science, [2]Stanford Education, [3]Stanford Psychology

April 25, 2026

Abstract

Measurement and learning are deeply intertwined in both human assessment and AI evaluation, yet they have largely been studied in isolation. In educational settings, assessments designed to measure a student's ability also influence their learning trajectory; in AI evaluation, benchmark performance shifts as models are updated—whether through routine training, deliberate fine-tuning, or incidental exposure to evaluation data. Both domains face a shared challenge: how to reliably measure a capability that changes over time, potentially in response to the measurement process itself. This need is increasingly urgent: in AI-tutored classrooms, it is difficult to tell whether a student is genuinely learning or the system is merely appearing helpful; in AI evaluation, benchmark scores conflate genuine capability with memorization of test items. Underlying these challenges is a structural parallel that has gone largely unexploited: at their core, both psychometric testing and AI benchmarking infer latent ability from behavioral observations of how subjects respond to items. Yet the two fields have developed complementary strengths—psychometrics brings over a century of construct theory and principled measurement frameworks, while machine learning brings powerful predictive tools that can extract signal from large, temporally structured data. This paper presents a roadmap that bridges these strengths through a unified framework organized around two predictive-measurement regimes—*static measurement*, where ability is inferred at a fixed point in time, and *dynamic measurement*, where ability is inferred as a trajectory. For each regime, we formalize the problem, survey representative approaches from both fields, and identify open problems. We then map human assessment and AI evaluation scenarios onto the framework in parallel, showing that the two domains face formally analogous challenges and identifying where solutions developed in one domain can inform progress in the other. By providing a shared language across these fields, this roadmap aims to accelerate progress on the common problem of measuring intelligent systems.

1 Introduction

Both psychometric testing and AI benchmarking infer ability from item responses. Both were designed around the assumption that ability is fixed during evaluation. In both fields, that assumption is breaking down. Psychometrics has mature tools for static measurement—construct theory, item response theory, validity frameworks—while AI benchmarking has large-scale response data and fast iteration cycles, though aggregate benchmark scores remain sensitive to prompt format, evaluation protocol, and data contamination. The subjects in both domains, however, are increasingly dynamic: students develop over the course of instruction, and AI systems change through fine-tuning, in-context learning, and continual updates. In the dynamic case, a single ability estimate is stale by the time it is reported, and the measurement problem is to recover a trajectory—potentially one whose evolution is shaped by the measurement process itself, through test practice and feedback on the human side or through in-context exposure on the AI side. Together, these cases define two measurement regimes—one where ability is treated as fixed, one where it is modeled as a trajectory—that organize the remainder of this roadmap.

This need is increasingly urgent. In *human assessment*, AI-tutored classrooms pose a basic question that current tools cannot answer: *is the student actually learning, or is the AI system simply getting better at appearing helpful?* In *AI evaluation*, benchmark scores conflate genuine capability with memorization of test items, and measured performance shifts dramatically depending on the evaluation protocol. In both cases, the core challenge is the same: disentangling genuine ability change from artifacts of the measurement process. Without reliable measurement, we cannot tell whether AI-enhanced education is working; equally, we cannot determine whether AI systems are genuinely advancing or merely overfitting to the benchmarks we use to track progress. We elaborate on these challenges in Section 2.

Despite this shared structure, the parallel between psychometric testing and AI benchmarking has gone largely unexploited. A standardized test records how students answer questions; an AI benchmark records how models perform on tasks. In the simplest static case, these observations form a response matrix of subjects $\times$ items, connecting measurement science to matrix completion and collaborative filtering; more generally, the data form temporally structured sequences of subject–item interactions. Yet the two fields have developed complementary strengths and weaknesses. *Psychometrics* brings over a century of construct theory—principled frameworks for defining what is being measured, establishing whether instruments measure it validly, and ensuring that measurements generalize across populations—but its predictive models remain relatively simple, and longitudinal data at the human time scale are expensive to collect. *Machine learning* brings powerful predictive tools—deep collaborative filtering, neural sequence models, and scalable Bayesian inference—that can extract signal from large, sparse, and temporally structured data, but AI evaluation lacks the deep construct-theoretic foundations that psychometrics provides. A key shared frontier is the *temporal* problem: tracking how ability evolves over time, in response to interventions, and potentially in response to the measurement process itself. In AI, this evolution is fast—models are updated weekly or continuously—generating rich temporal data but outpacing current evaluation methodology. In human assessment, longitudinal data are sparse and expensive, but the underlying theory of learning dynamics is far more developed. This roadmap aims to bridge these complementary strengths, providing a unified framework through which tools and insights can transfer in both directions (Figure 1).

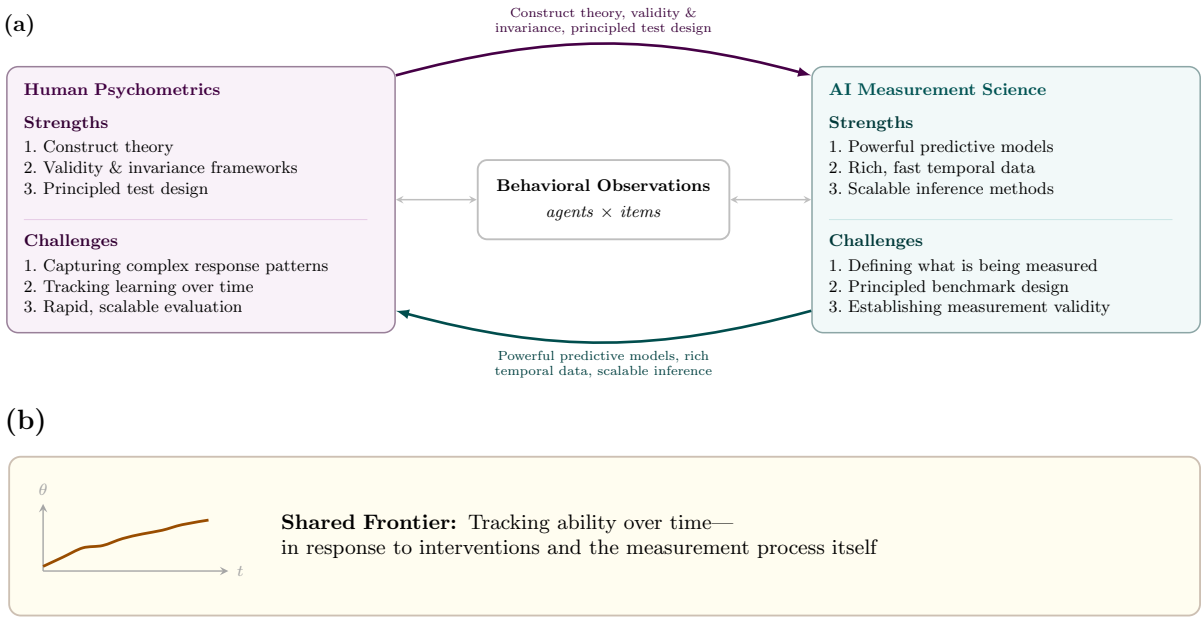

Figure 1: **(a)** Psychometrics and AI evaluation address the same abstract problem—inferring latent ability from behavioral observations of how agents respond to items—but bring complementary strengths, forming a virtuous cycle: psychometrics contributes construct theory and principled measurement design, while machine learning offers powerful predictive models and scalable methods for temporally structured data. **(b)** The shared frontier is the *temporal* problem: modeling how ability evolves over time, which neither field has fully solved. This roadmap bridges the two, enabling tools and insights to transfer in both directions.

Existing frameworks in both domains reflect this divide. In *human assessment*, classical test theory (CTT) [Spearman, 1904, Edgeworth, 1888] and item response theory (IRT) [Lawley, 1943, Rasch, 1993]

estimate ability from data collected over a short time span, as in standardized tests like the TOEFL [Way and Reese, 1990]. Separately, training paradigms such as adaptive tutorial systems (e.g., Duolingo [Munday, 2016], Khan Academy) focus on adapting instruction to improve performance. Attempts to bridge measurement and learning remain partial: Bayesian knowledge tracing [Corbett and Anderson, 1994] models learning dynamics but uses discrete mastery states rather than continuous latent traits, limiting its connection to psychometric measurement; deep knowledge tracing [Piech et al., 2015] and its successors improve predictive accuracy but largely abandon the interpretable latent structure that makes psychometric guarantees (e.g., item-invariant ability estimates) possible. In *AI evaluation*, a parallel gap exists: Elo-based rating systems [Boubdir et al., 2023] track relative model performance over time but lack a principled account of what is being measured or how measurement interacts with model updates; static benchmarks provide snapshots of capability but do not account for how models change between or during evaluations. What is missing in both domains is not another model, but a *roadmap* that situates these approaches within a common space, clarifies the assumptions each makes about dynamics and feedback, and identifies which problems remain open.

This roadmap addresses two guiding questions: (1) How should we measure an ability that changes over time—and potentially in response to the measurement process itself? (2) To what extent do challenges in AI evaluation (benchmark contamination, in-context learning, post-training shifts) mirror well-studied problems in human psychometrics, and can solutions transfer across domains?

Scope. This paper is a *roadmap*, not a new model or algorithm. It provides shared language and a framework that organizes existing work and identifies open problems. We focus on ability measurement through item-response interactions—the setting common to psychometric testing and AI benchmarking. We do not address safety evaluation, red-teaming, or adversarial robustness testing: these involve normative judgments where the "correct" response depends on context and values, posing construct validity questions that are qualitatively different from ability estimation and would require separate treatment. We discuss how the framework extends to non-binary response formats (pairwise preferences, graded responses, multi-turn interactions) in Section 4. Throughout, we use *subject* to refer to the entity being measured—whether a human learner or an AI system—and *evaluator* for the entity administering items and making decisions. Sections 4–5 develop the framework and illustrate each setting with concrete instantiations from both human assessment and AI evaluation. When examples are domain-specific, we indicate whether they pertain to *human assessment*, *AI evaluation*, or both.

Outline. We begin with two grand challenges that ground the abstract framework in concrete, testable goals (Section 2). We then review classical measurement and learning models—item response theory and knowledge tracing—alongside the AI evaluation literature (Section 3). Next, we introduce notation and assumptions for the interaction between a subject and an evaluator (Section 4). Building on this framework, we develop two predictive-measurement regimes—*static measurement* (inferring a scalar ability θ_i) and *dynamic measurement* (inferring a trajectory $\theta_i(t)$; Section 5)—and identify open research directions for each. For each regime, we illustrate the formal problem with concrete instantiations from both human assessment and AI evaluation, showing that the two domains face formally analogous challenges—and noting where the parallels are tight and where the underlying mechanisms diverge.

2 Grand Challenges

Before introducing the formal framework, we present two grand challenges that ground the abstract framework in concrete, testable goals. Like the Characters Challenge and Frostbite Challenge in Lake et al. [2017], these are not merely open problems but *organizing benchmarks*: each defines a measurable objective that, if achieved, would demonstrate that the unified measurement science advocated in this paper has practical force. The two challenges are deliberately complementary. The first asks whether we can measure a subject's ability at a fixed point in time from partial observations; the second asks whether we can model and predict how that ability *evolves* over time. Together, they correspond to the two measurement regimes—static and dynamic—that organize the remainder of this paper, and the second builds on the first: reliable trajectory prediction requires reliable point-in-time estimation as its foundation.

Crucially, both challenges apply equally to human learners and AI systems. Whether the subjects are students taking a math test or language models answering benchmark questions, the data have the same structure: a table of subjects by items, recording who got what right—and, in the dynamic case,

when. The core inference problems—predicting missing entries, forecasting trajectories, separating genuine change from observation artifacts—are formally identical across domains. This structural parallel has been articulated most comprehensively by Hernandez-Orallo [2017], who argues for a "universal psychometrics" that applies the same measurement-theoretic tools to any intelligent agent—human, animal, or machine. The two domains have complementary gaps. Human assessment has decades of methodological infrastructure—item response theory, differential item functioning, computerized adaptive testing, growth-curve modeling—refined on well-curated longitudinal datasets, but prospective validation of predictions is rare and data scale is limited. AI evaluation has unprecedented data scale and fast iteration cycles, but benchmarks are often released without defined constructs [Bean et al., 2025] and without the psychometric tooling that would make them trustworthy instruments. A grand challenge creates the conditions for these complementary strengths to meet.

Challenge 1: Predictive Modeling for Static Measurement. *Problem.* Even when a subject's ability can be treated as fixed, measuring it well remains hard. Evaluation today is retrospective and sparse: it tells us how a model performed on yesterday's benchmark or how a student scored on last month's test, but not how either would perform on new items drawn from the same construct, nor how to extrapolate from one benchmark to another. In other scientific fields—meteorology, epidemiology, economics—forecasting under uncertainty is foundational infrastructure. In AI evaluation, it is largely absent. Benchmarks increasingly function as optimization targets rather than scientific instruments [Ganguli et al., 2023]. A recent systematic review of 445 LLM benchmarks found that roughly 22% are published without even defining what they measure, and most lack the statistical tests for construct validity that psychometrics would consider routine [Bean et al., 2025]. The result is a structural gap: we lack a science of *predictive* evaluation even in the simplest, stationary setting.

The consequences of this gap are not abstract. Consider GSM8K, one of the most widely used benchmarks for mathematical reasoning. Approximately 8.8% of its items contain errors—ambiguous wording, incorrect answer keys, or grading failures as elementary as marking "$7.00" incorrect while accepting "$7" for the same question [Truong et al., 2025a]. These errors are invisible to aggregate accuracy scores. But when psychometric diagnostics flag and remove the problematic items, model rankings change dramatically: on the revised benchmark, DeepSeek-R1 rises from third-lowest to second place overall (Figure 2). The leaderboard that the field relies on to allocate research effort and deployment decisions is, in part, an artifact of measurement error in the items themselves. If we cannot even trust retrospective evaluation on a fixed benchmark, the prospect of predictive evaluation—forecasting performance on unseen items and unseen subjects—may seem remote. Yet as we argue below, the same measurement-theoretic tools that expose these problems also provide the foundation for solving them.

The same gap exists, in milder form, in human assessment. A student's score on a standardized test is treated as a point estimate of ability, but the question that matters for educational decisions—how will this student perform on a different test drawn from the same construct, or on items that have not yet been pilot-tested?—requires prediction, not just measurement. Psychometrics provides the formal machinery for such predictions, but this predictive capacity is rarely validated prospectively and breaks down when the assumption of a fixed, unidimensional ability is violated.

The challenge. The core data structure in both human assessment and AI evaluation is a response table: rows are subjects (students or models), columns are items (test questions or benchmark tasks), and each entry records whether the subject answered the item correctly. Most of this table is missing—no student takes every test, and no model is evaluated on every benchmark. The static challenge is to predict the missing entries, under progressively harder conditions.

Level 1: Filling in the gaps. The simplest version asks: given a partially filled response table, can we predict the missing entries? This is the setting where psychometrics has the longest track record. Computerized adaptive testing (CAT)—the workhorse of large-scale human assessment for decades—does exactly this: it selects questions sequentially based on the student's estimated ability, achieving the same measurement precision as a full-length test with far fewer items. The GRE, GMAT, and many state assessments have used this approach for years [Scalise and Gifford, 2006].

On the human side, ROAR-CAT demonstrates these efficiency gains concretely: an IRT-based adaptive reading assessment achieves 40% efficiency gains over non-adaptive administration, reaching reliability of 0.9 with 75 adaptively selected items versus 125 random items [Ma et al., 2025]. Recent work shows that the same approach transfers directly to AI evaluation. Polo et al. [2024] demonstrate that IRT-based item selection can reduce benchmark size by an order of magnitude while preserving ranking accuracy, and Truong et al. [2025b] extend this with adaptive item selection across 22 benchmarks and 172 language models, recovering benchmark-level accuracy within 2% error using only 1–18% of items. The efficiency

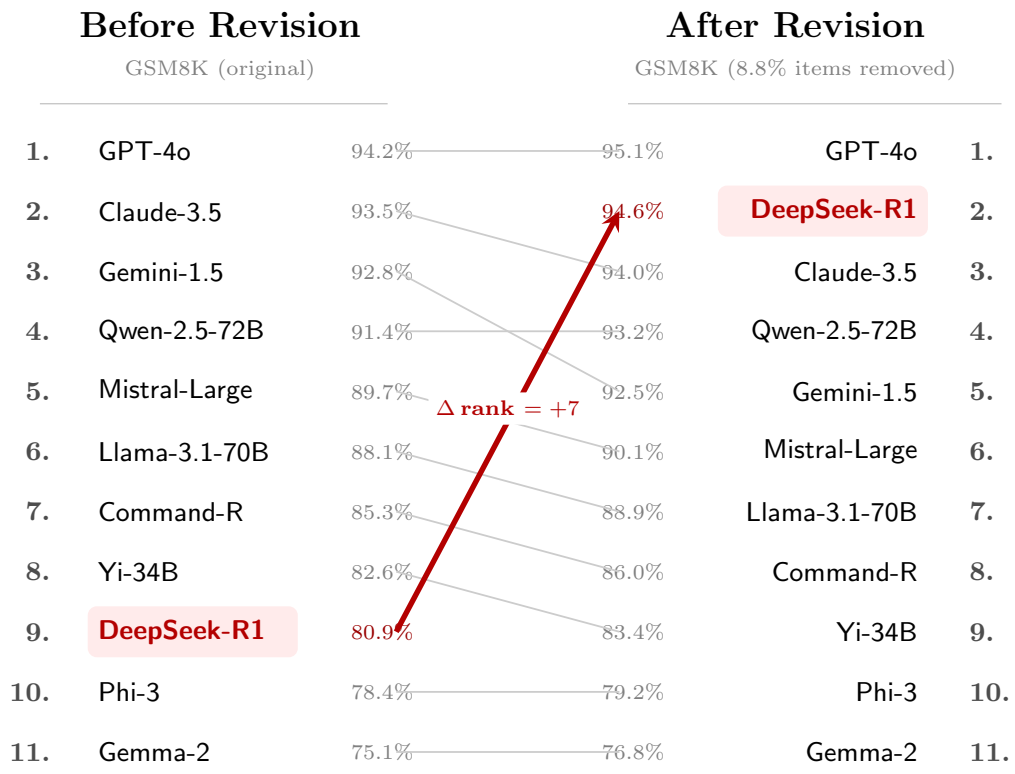

Figure 2: The cost of measurement error in benchmarks. Model rankings on GSM8K before (left) and after (right) removing approximately 8.8% of items flagged as invalid by psychometric diagnostics [Truong et al., 2025a]. DeepSeek-R1 rises from third-lowest (rank 9) to second place (rank 2), a shift of seven positions. The leaderboard that guides research investment and deployment decisions is, in part, an artifact of flawed benchmark items.

gains can be dramatic: on AIR-Bench, a safety benchmark with 4,985 items, adaptive selection achieves 95% reliability from just 31 items—a 99.4% reduction in evaluation cost (Figure 3). These are instances of the same CAT pipeline used to administer the GRE to millions of human examinees, transferred to AI evaluation with learned difficulty predictors in place of empirical item calibration [Zhuang et al., 2023].

This level of the challenge is *largely solved* in both domains. It serves as the baseline—and as proof that cross-pollination works.

Level 2: Predicting difficulty for never-seen items. A harder version asks: given a brand-new item that has never been administered to anyone, can we predict how hard it will be? If so, we can evaluate subjects on fresh items without the expensive pilot studies that psychometrics traditionally requires. Embretson's *cognitive design system* approach [Embretson, 1998] provides the theoretical foundation: by identifying the cognitive features that determine an item's difficulty (number of reasoning steps, degree of abstraction, vocabulary level), one can predict item parameters from task structure rather than requiring empirical calibration. The explanatory IRT framework [De Boeck and Wilson, 2004] provides the statistical machinery for embedding such features as covariates. Recent work has begun to realize this vision empirically. On the human side, difficulty of reading comprehension items can be predicted from linguistic and contextual features, achieving correlations of $r = 0.77$ with empirically calibrated parameters across standardized tests in grades 3–8 [Kapoor et al., 2025]. On the AI side, learned models that map benchmark items to difficulty parameters enable adaptive testing on items that no model has ever seen (see Level 1 above).

This level is *emerging*—feasible in restricted domains but not yet general. Predicting difficulty from item features works well for reading comprehension and multiple-choice factual questions, but remains untested for more complex task types (multi-step reasoning, open-ended generation, interactive tasks).

Level 3: Simulating new subjects. A still harder version asks: given a subject that has never been observed—a student who has not yet enrolled, or a model that has not yet been released—can we generate realistic predictions of how they would respond? This amounts to synthesizing entire rows of the response table. Early results are promising but uneven. Language models fine-tuned to simulate student responses

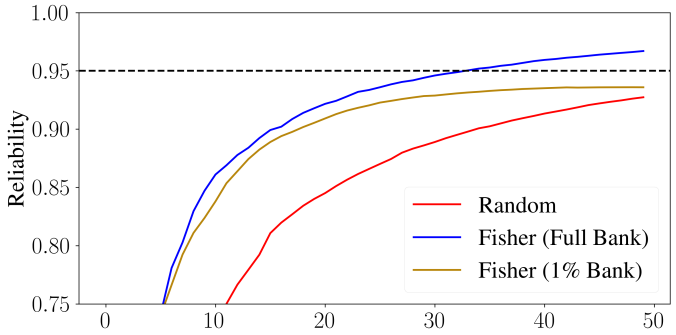

Figure 3: Adaptive item selection for efficient AI evaluation, illustrated on AIR-Bench (4,985 safety items). Fisher information–based adaptive selection from the full item bank (blue) reaches 95% reliability in ability estimation from approximately 31 items, while random selection (red) requires the full 50-item budget without reaching the threshold. Adaptive selection from a restricted 1% bank (gold) shows that item bank diversity, not just selection strategy, is critical. The same IRT + CAT pipeline has been used in human standardized testing (GRE, GMAT) for decades.

to reading efficiency items can produce synthetic test forms whose psychometric properties correlate at $r = 0.93$ with expert-developed instruments validated on 234 real students [Zelikman et al., 2023]. But for richer behavioral domains the picture is less encouraging: when language models are prompted to imitate programming students solving C++ problems, the best model achieves ability parameter recovery of only $\rho \approx 0.36$—substantially better than chance, but far from the fidelity needed for reliable calibration [Truong et al., 2025c].

This level is at the *frontier*. The gap between $r = 0.93$ for simulating reading responses and $\rho \approx 0.36$ for simulating programming trajectories highlights a fundamental tension: models trained predominantly on correct code struggle to reproduce the novice error patterns that characterize real student learning. Current language models cannot simultaneously serve as expert problem-solvers and realistic student simulators—a limitation with direct implications for using synthetic data to calibrate assessments.

Why a grand challenge? Grand challenges catalyze new scientific fields. ImageNet [Deng et al., 2009] did this for representation learning; the Netflix Prize [Bennett and Lanning, 2007] did this for collaborative filtering. A predictive evaluation challenge can do the same for AI measurement science. The central obstacle is not the absence of methods but the absence of *ground truth*: there is currently no infrastructure for testing whether a predictive evaluation method works. Each cycle of the challenge generates a new layer of ground truth—real forecasts, real holdouts, real outcomes—accumulating into the empirical foundation that predictive evaluation science cannot build any other way. Shared data resources like the Item Response Warehouse, which aggregates over 900 standardized item response datasets from human assessments [Domingue et al., 2025], provide a model for the kind of open infrastructure that predictive evaluation science requires. On the human side, the same infrastructure would enable prospective validation of psychometric predictions in educational settings, where such validation is surprisingly rare despite the maturity of the methodology.

Challenge 2: Predictive Modeling for Dynamic Measurement. *Problem.* The static framing of Challenge 1 is a useful fiction. In reality, subjects change. A student's ability shifts with instruction, practice, and forgetting; a model's capability shifts with fine-tuning, data-mixture updates, scaffolding changes, and deployment drift. A snapshot measurement at a single point in time is stale by the time it is published. Yet current practice in AI evaluation treats models at release as fixed artifacts, rarely revisiting them, and even in standardized human assessment the machinery for modeling ability *as a trajectory* rather than a point estimate is deployed only in narrow contexts.

A concrete example illustrates the cost of treating evaluation as a one-shot snapshot. When GPT-3.5's safety performance is tracked across successive versions using IRT-based ability estimation, the trajectory is non-monotonic: estimated safety ability rises steadily from January 2023 to a peak in June 2023, then drops sharply with the November 2023 update (Figure 4). A single evaluation at any one time point would have missed the regression entirely. A model deployed on the basis of its June safety score continued operating through a substantial, undetected decline—not because the data were unavailable, but because the measurement framework in widespread use does not represent ability as a time-varying quantity at all.

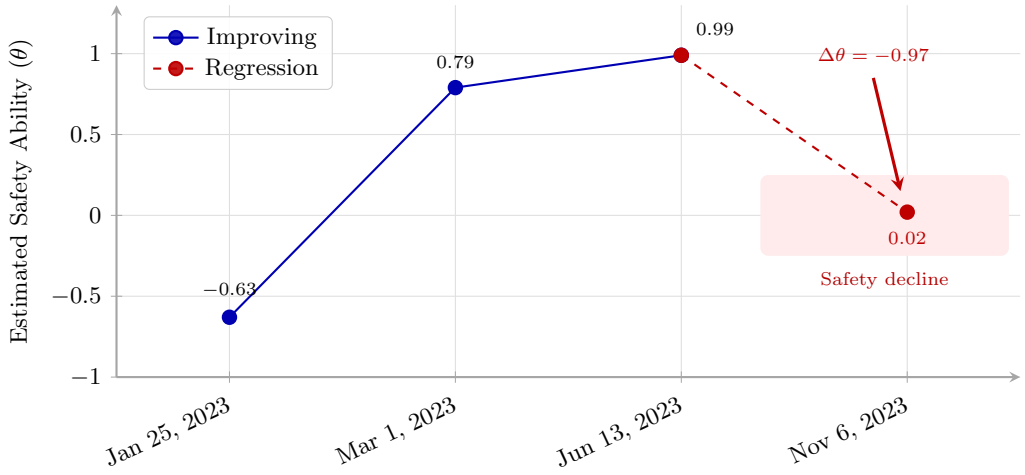

Figure 4: The cost of snapshot evaluation. Estimated safety ability (θ) of GPT-3.5 across successive versions, measured using IRT-based ability estimation on a safety benchmark. Safety ability rises steadily from January to June 2023, then drops sharply with the November 2023 update ($\Delta\theta = -0.97$). A single evaluation at any one time point would miss the regression entirely; continuous, trajectory-aware measurement is needed to detect and characterize such declines.

The same issue arises in education, with a subtler twist. A school district deploys an AI tutoring system; at year end, standardized test scores improve. But the tutor's own internal assessments of student progress may be unreliable—because it adapted item difficulty to each student throughout the year, high accuracy on tutor-selected items could reflect genuinely improved ability, simply easier items, or a tutor that learned to optimize for engagement rather than learning. The result is a measurement problem that the static response-table framing cannot express: the items asked at time t depend on what the subject answered at times $t' < t$, and the observation process is entangled with the very state it is trying to measure. Disentangling genuine change from measurement artifact requires jointly modeling the trajectory and the adaptive observation policy—a problem that neither classical psychometrics nor standard machine learning has addressed in isolation.

The challenge. The data structure now extends from a two-dimensional response table (subjects × items) to a three-dimensional one (subjects × items × time). Ability θ is no longer a fixed quantity to be inferred; it is a trajectory $\theta(t)$ to be modeled, typically governed by latent dynamics that mix slow drift, discrete interventions (a major fine-tuning run, a semester boundary), and within-session learning from feedback. The challenge is to predict this trajectory, under progressively harder conditions.

Level 1: Reconstruction. The simplest version asks: given dense longitudinal observations, can we recover the trajectory $\{\theta(t)\}$? This is the dynamic analog of Challenge 1 Level 1—monitoring rather than one-shot measurement. On the human side, growth-curve modeling and dynamic IRT have a long track record: Vandemeulebroecke et al. [2017] model semi-annual neuropsychiatric assessments as subject-specific linear drift; Martin and Quinn [2002] track Supreme Court ideal points using a Gaussian random walk; Abbakumov et al. [2019] extend the Rasch model to MOOCs, decomposing between-session drift from within-session learning; and a line of work on adaptive learning platforms [Klinkenberg et al., 2011, Kadengye et al., 2014, Bolsinova et al., 2022] deploys dynamic latent variable models at scale. Simpler update rules such as Elo [Pelánek, 2016] provide a computationally cheap alternative when the dynamics are smooth.

On the AI side, the infrastructure exists but is almost never used. Model versions are routinely released without cross-calibration to their predecessors, benchmarks are not rerun against old models, and no standard pipeline treats θ as time-varying. The GPT-3.5 safety trajectory in Figure 4 illustrates both that the measurement is tractable and that it is essentially never performed. This level of the challenge is *largely solved in principle* on the human side and *largely undeployed* on the AI side.

Level 2: Forecasting. A harder version asks: given the trajectory so far, can we predict how the subject will perform in the future—a student six months from now, a model after its next training run, a new release on a benchmark that did not exist at the time of the last evaluation? This is the dynamic analog of cold-start: extrapolation along the *time* axis rather than the subject or item axis. Existing approaches are partial. AI scaling laws forecast aggregate loss as a function of compute and data, but not

item-level behavior, and they break down across architecture changes. Dynamic IRT and growth-curve models capture gradual drift but cannot anticipate discrete interventions like a data-mixture change or the introduction of a new system prompt. Human longitudinal studies demonstrate that forecasting under assumptions of stationary dynamics is unreliable when regime changes occur [Cepeda et al., 2009]. The hardest version combines forecasting with cold-start along the other axes: a model that has not yet been released, evaluated on a benchmark that has not yet been constructed. No existing method credibly achieves this, and—as with static Level 3—the infrastructure needed to even *test* such methods does not yet exist: there are no held-out prospective datasets, no mechanisms for comparing forecasts against outcomes that did not exist when the forecast was made. Creating this infrastructure is itself a primary contribution of the challenge.

This level is at the *frontier* for both domains.

Level 3: Disentangling change from observation artifact. The hardest version concerns identifiability. When observations are not exogenous—when the items asked depend on the subject's prior responses, or when training data is generated on-policy—the observation process is entangled with the state it is trying to measure. The school-district example makes this concrete on the human side: an adaptive tutor's response patterns mix student ability with item-selection policy, and pre–post gains on tutor-selected items can reflect either. On the AI side, an analogous confound arises in reinforcement learning from on-policy rollouts and in benchmark contamination: the evaluation data is generated by, or has been seen by, the very system being evaluated. Position effects—where the same item becomes easier or harder depending on where it appears in the evaluation sequence—have been formally modeled in human assessment through position-sensitive IRT [Kanopka and Domingue, 2025]; analogous prompt-ordering effects in LLM evaluation are documented [Lu et al., 2022] but rarely modeled. When $\mathcal{L}$ (within-session learning) and $\mathcal{S}$ (between-session drift) are both present, the two processes may not be separately identifiable from response data alone (see Section 4). Recovering the underlying trajectory in the presence of these feedback loops is an open problem in both domains.

This level is at the *frontier*, and it is where the two domains have the most to offer each other: psychometrics has formal tools for identifiability and position effects that AI evaluation lacks; AI evaluation has orders-of-magnitude more data per subject, which is precisely what identifiability arguments require.

Why a grand challenge? The same logic that motivates the static challenge applies with greater force here. Static measurement errors can be caught eventually through replication and reanalysis; dynamic measurement failures—an undetected safety regression, a student whose trajectory is masked by adaptive item selection—compound over time and become harder to recover from the longer they go unmeasured. Yet dynamic measurement suffers from an even sharper ground-truth problem than static: prospective trajectory predictions are essentially never recorded, let alone scored. A grand challenge that demands registered forecasts of future model performance, scored against the trajectories those models actually follow, would create the ground-truth record that a predictive science of dynamic measurement requires—and would do so in exactly the regime (frontier models, fast iteration cycles) where the need is most acute.

These two challenges correspond to the two measurement regimes—static and dynamic—that organize the remainder of this paper. Challenge 1 maps to the static measurement regime of Section 5; Challenge 2 maps to the dynamic measurement regime, with its distinction between between-session drift and within-session learning. The framework developed below provides the formal tools for reasoning about both: a unified model (Section 4) that captures how ability evolves through feedback and spacing, and a setting-by-setting treatment (Section 5) that organizes the space of measurement scenarios. Throughout, we use the two challenges as touchstones, returning to them when illustrating how specific technical advances bring us closer to the concrete goals they define.

3 Related Work

This paper draws on several research traditions that have developed largely in parallel: psychometric measurement, dynamic ability modeling, learning and training, and AI evaluation. We briefly survey each and identify the gaps that motivate the unified framework in Section 4.

Static measurement and item response theory. Item response theory (IRT) addresses a core limitation of classical test theory—that ability estimates depend on which items were administered—by modeling both the ability of the subject and the properties of items such that, under certain assumptions, the parameters exhibit *invariance*: item parameters do not depend on which subjects were sampled, and

ability parameters do not depend on which items were administered [Hambleton et al., 1991]. This invariance property enables *adaptive testing*: because item parameters calibrated on one sample remain valid for new subjects, items can be pre-calibrated and then selectively administered based on performance [Van der Linden et al., 2000]. Crucially, invariance is an *aspiration* that holds under correct model specification, not a guaranteed property of IRT. When the model is misspecified—for example, when items are truly multidimensional but a unidimensional model is fit, or when responses exhibit local dependence (i.e., the response to one item directly influences the response to another, violating the conditional independence assumption)—invariance breaks down and ability estimates become biased [Embretson and Reise, 2000]. *Differential item functioning* (DIF) formalizes one important mode of failure: an item exhibits DIF when subjects with the same underlying ability but from different subpopulations have systematically different probabilities of answering correctly, indicating that the item measures something beyond the intended construct [Holland and Wainer, 1993]. DIF is especially relevant when extending IRT to AI evaluation, where the "population" of models is far more heterogeneous than a typical human examinee population—models differ in architecture, training data, and optimization procedure, any of which may cause an item to function differently across model families even after controlling for overall ability. In the canonical formulation, a subject with ability θ answers items with parameter z_t, and the probability of a correct response is given by the 1-parameter logistic (Rasch) model: $p(y_t = 1|\theta, z_t) = \sigma(\theta - z_t)$, where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function and z_t is the item difficulty. More complex models add discrimination (a_t) and guessing (c_t) parameters—the 2PL and 3PL models, respectively—which are particularly relevant for AI benchmarks, where some items may fail to discriminate between models or where guessing behavior differs systematically from human test-takers. Parameters are estimated via expectation-maximization or MCMC. Beyond psychometrics, IRT has been applied to neuropsychiatric monitoring [Vandemeulebroecke et al., 2017], political ideology measurement [Martin and Quinn, 2002], and marketing research [De Jong et al., 2008]. We provide additional historical context on classical test theory and computerized adaptive testing in Appendix C.

Dynamic measurement. The foundational IRT framework assumes a static ability estimated within a single session, but a substantial body of work has relaxed this assumption. Growth models and change-score methods study ability trajectories across time points [Embretson, 1991]; dynamic factor analysis and state-space models treat ability as a latent state evolving according to a transition model [Molenaar, 1985, Zhang and Li, 2022]; and longitudinal student modeling in learning analytics tracks learner progress over time [Lang et al., 2017]. The interim and formative assessment literature is particularly extensive: programs that evaluate students across shorter assessments throughout the year are now widespread in practice [Modan, 2023, Andrade and Cizek, 2010]. Despite this rich body of work, a gap remains: most dynamic extensions either sacrifice the psychometric guarantees of classical IRT (e.g., item-invariant ability estimates) or treat each assessment occasion independently rather than modeling the full trajectory. This gap motivates the framework we develop in Section 4.

Training and learning. When the goal shifts from measuring ability to improving it, knowledge tracing (KT) becomes the dominant modeling paradigm. Bayesian Knowledge Tracing (BKT) models each skill as a hidden Markov model with binary states (mastered or not), parameterized by guess, slip, and learn probabilities [Abdelrahman et al., 2023]; these representations are then used to select items that maximize learning [Piech et al., 2015, Rafferty et al., 2016]. IRT and BKT differ in a fundamental structural way: IRT models ability as a *continuous* latent trait $\theta \in \mathbb{R}$, while BKT models knowledge as a *discrete* binary state. Continuous traits support graded measurement and the psychometric guarantees that underpin high-stakes assessment; discrete mastery states are better suited to curriculum sequencing but do not provide the same measurement properties. *Cognitive diagnostic models* (CDMs) [Rupp et al., 2010] occupy a middle ground: they model mastery as a binary vector over multiple fine-grained skills (via a Q-matrix that maps items to required skills), providing richer diagnostic profiles than a single θ while retaining a structured latent representation. CDMs are relevant to AI capability evaluation, where one might ask not just "how able is this model?" but "which specific skills has it mastered?"—though the number of skills and the Q-matrix must be specified in advance, which is nontrivial for open-ended AI benchmarks. Deep Knowledge Tracing (DKT) [Piech et al., 2015] and its successors—including attention-based [Ghosh et al., 2020] and transformer architectures [Choi et al., 2020]—replace explicit latent structure with learned neural representations, achieving strong predictive accuracy but sacrificing interpretability and calibrated uncertainty. Bridging the predictive power of deep models with the interpretable latent structure of IRT remains an important open direction. See Appendix C for the full BKT formulation.

AI evaluation. The dominant paradigm for evaluating AI systems mirrors the static measurement regime of our framework: a model is evaluated once on a fixed benchmark (e.g., MMLU, HELM, GPQA), producing accuracy scores that serve as point estimates of ability. There are growing efforts to apply psychometric methods to this setting: Martinez-Plumed et al. [2019] apply IRT to machine learning algorithm evaluation, and Polo et al. [2024] show that IRT-based item selection can reduce benchmark size by an order of magnitude while preserving ranking accuracy. Elo-based rating systems such as Chatbot Arena [Zheng et al., 2024, Boubdir et al., 2023] track relative performance through pairwise comparisons; the underlying Bradley–Terry model [Bradley and Terry, 1952] is formally equivalent to the Rasch model with the opponent's ability replacing item difficulty, providing a natural bridge between pairwise preference evaluation and the IRT framework—though current deployments do not fully exploit this connection for uncertainty quantification or adaptive matchup selection. Beyond static benchmarks and pairwise rankings, AI evaluation is increasingly moving toward open-ended generation tasks scored by LLM-as-judge systems [Zheng et al., 2024, Shankar et al., 2024] and multi-step agentic tasks such as SWE-bench [Jimenez et al., 2024], where the notion of a discrete "item" with a binary outcome becomes ambiguous. These newer evaluation paradigms raise measurement questions—scorer reliability, construct definition, item boundary specification—that the IRT framework can address through extensions such as polytomous response models [Samejima, 1969] and rater-effect modeling within generalizability theory, as we discuss in Section 4. Meanwhile, challenges familiar to psychometrics are emerging across all forms of AI evaluation: benchmark contamination compromises measurement validity when evaluation data appears in training sets [Ganguli et al., 2023], sensitivity to prompt ordering and few-shot examples mirrors item-order effects in human assessment [Lu et al., 2022], benchmark saturation drives the creation of progressively harder evaluations [Glazer et al., 2024] in a dynamic reminiscent of ceiling effects in psychometric testing, and the question of whether benchmarks measure coherent constructs parallels longstanding concerns about construct validity in testing [Raji et al., 2021]. These challenges map directly onto the framework developed in Section 5.

Positioning of this paper. Recent work has begun applying specific psychometric tools to AI evaluation. Burnell et al. [2023] propose transitioning from task-oriented benchmarking to construct-oriented evaluation using IRT and validity evidence; Zhuang et al. [2023] demonstrate efficiency gains from computerized adaptive testing applied to language models; Zhou et al. [2025] apply a 4-parameter IRT model to 11 benchmarks, revealing significant variation in item quality—though notably, this work assumes unidimensionality throughout without testing whether each benchmark measures a single coherent construct; and Ilic and Gignac [2024] apply confirmatory factor analysis to 591 language models, finding a strong general ability factor that accounts for 66% of variance across benchmarks, echoing the positive manifold observed in human cognitive testing. These contributions apply the tools of static measurement (no feedback) to a new domain, but have not yet engaged with the deeper psychometric toolkit—generalizability studies, differential item functioning, multidimensional IRT—that would be needed to establish whether current benchmarks satisfy the assumptions these methods require.

Our paper differs in three ways. First, we provide a *framework spanning static and dynamic measurement*, distinguishing a scalar inference target θ_i from a trajectory target $\theta_i(t)$ and organizing open problems within each; existing work addresses only the static regime. Second, we *formally model the interplay between measurement and learning*—how feedback during evaluation changes the quantity being measured, and how learning and stochastic drift jointly shape ability trajectories—a problem that neither the psychometrics-for-AI literature nor the knowledge tracing literature has addressed in full generality. Third, we *draw parallel application maps* across human assessment and AI evaluation within each regime (Section 5), showing that the two domains face formally analogous challenges at every level—while noting where the parallels are tight and where they are looser—rather than treating psychometrics as a toolkit to be imported into AI evaluation.

4 Preliminaries

We introduce the notation and assumptions used throughout the paper. The question that organizes the framework is whether the subject's ability changes during or between evaluations. If ability is static, a single measurement session suffices. If ability is dynamic, we partition all sources of change along a temporal boundary: *within-session change* ($\mathcal{L}$), encompassing any ability change that occurs while the subject is being evaluated (feedback-driven learning, practice effects, fatigue, strategic adaptation), and *between-session change* ($\mathcal{S}$), encompassing any ability change that occurs during gaps in evaluation

(self-study, forgetting, continued training, developmental growth). These two categories are exhaustive by construction and can occur alone or together.

This distinction yields two problem regimes, visualized in Figure 5. When the subject's ability can be treated as fixed during evaluation, we are in *static measurement*: the inference target is a scalar θ_i per subject, and the open problems concern generalizing across missing entries, unseen items, and unseen subjects. When ability changes—via between-session drift ($\mathcal{S}$), within-session learning ($\mathcal{L}$), or both—we are in *dynamic measurement*: the inference target is a trajectory $\theta_i(t)$, and the open problems concern reconstruction, forecasting, and identifiability under endogenous observation. The two regimes share the same response-table substrate but have different inference targets and different native difficulties; neither reduces to the other.

Connection to summative and formative assessment. The two regimes above parallel a well-established distinction in educational assessment between *summative* and *formative* assessment [Black and Wiliam, 1998, Scriven, 1967]. Summative assessment—evaluation administered at the end of a unit of instruction to certify what a subject has learned—corresponds directly to *static measurement*: ability is treated as fixed, no feedback is provided, and the sole goal is to produce an accurate point estimate. Formative assessment—evaluation embedded within instruction so that subsequent items or activities can be chosen adaptively—falls within *dynamic measurement*: the feedback may itself shift ability, and the evaluator must model a trajectory $\theta_i(t)$ rather than a fixed θ_i. Our framework makes this distinction explicit and supplies the formal tools—trajectory models, identifiability under endogenous observation, forecasting—that a science of formative assessment requires.

This mapping extends beyond human education. In AI evaluation, a static benchmark administered once to rank models resembles summative assessment, while an evaluation protocol that provides in-context examples or iterative feedback to probe a model's adaptive capabilities resembles formative assessment. The framework applies equally to both domains.

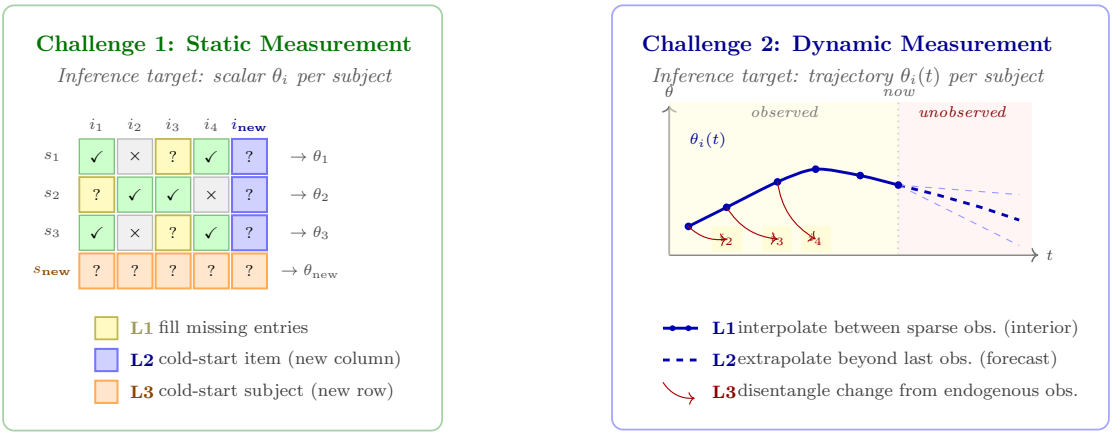

Figure 5: Two parallel predictive problems on a common response substrate. **Left (Challenge 1, static):** a subjects-×-items response table with observed entries (✓/×), missing entries to be filled (L1), a cold-start item column (L2), and a cold-start subject row (L3); the inference target is a scalar ability θ_i. **Right (Challenge 2, dynamic):** an ability trajectory $\theta_i(t)$ inferred from sparse longitudinal observations. The solid segment interpolates between observed points inside the observed range (L1); the dashed segment extrapolates past the last observation, with an uncertainty cone (L2). Red arrows indicate that the item selected at t_{k+1} depends on the response at t_k, entangling the observation process with the state being inferred (L3). The two challenges share structure but not difficulty: Challenge 1 is not a slice of Challenge 2—each has native inference problems (construct validity, subject simulation; identifiability under feedback loops) that do not reduce to the other.

An evaluator interacts with a subject with initial ability $\theta_0 \sim p(\theta_0)$. At time $t \in \mathbb{R}$, the evaluator administers an item q_t drawn from an item bank $\mathcal{Q}$ with associated parameter z_t (e.g., item difficulty in the Rasch model). The evaluator observes a binary outcome y_t (1 is correct). Within a session, the subject's ability may change as a function of the items encountered and the outcomes observed: $\theta_{t+1} = \mathcal{L}_\tau(\theta_t, z_t, y_t)$. This formalization captures feedback-driven learning as the primary mechanism, but also subsumes practice effects and the testing effect [Roediger and Karpicke, 2006], since the functional form depends on the item and response regardless of whether explicit feedback is provided.

Between sessions, the subject's ability may change due to processes unobserved by the evaluator (e.g., self-study, forgetting, continued pre-training, developmental growth). We model this as a stochastic process: $\theta_{t+1} = \mathcal{S}_\psi(\theta_t)$, where $\psi \sim p(\psi)$ are fixed, subject-specific, and learnable. We assume between-session change is smooth enough that ability cannot change too quickly. Together, $\mathcal{L}$ and $\mathcal{S}$ partition all sources of ability change along the session boundary: within-session change governs what happens during evaluation, and between-session change governs what happens between evaluations. Our notation is summarized in Table 2.

Identifiability of within-session and between-session change. When both $\mathcal{L}$ and $\mathcal{S}$ are present, the two processes may not be separately identifiable from response data alone. The fundamental difficulty is that the evaluator observes only a sequence of item responses—not the ability trajectory directly—and a given response pattern may be consistent with multiple decompositions of ability change into within-session and between-session components. Separate identification generally requires *structural constraints* from the experimental design. Three conditions are particularly relevant: (i) *no-feedback measurement probes*—interspersing items on which no feedback is provided within an otherwise feedback-rich session allows the evaluator to observe ability at points where $\mathcal{L}$ is inactive, anchoring the within-session trajectory; (ii) *session boundary variation*—if the same subject is observed under sessions of different lengths or with different inter-session gaps, the relative contributions of $\mathcal{L}$ and $\mathcal{S}$ vary while the underlying processes remain the same, providing leverage for identification; and (iii) *parametric smoothness assumptions*—imposing continuity or differentiability constraints on $\mathcal{S}$ (e.g., via a GP prior) restricts the set of admissible between-session trajectories, improving identifiability at the cost of model-dependence. When none of these conditions hold, estimates of $\mathcal{L}$ and $\mathcal{S}$ are confounded: a large apparent within-session gain may instead reflect unobserved between-session change, and vice versa. We revisit this challenge in the open problems of Section 5.

The exogeneity assumption. We assume the dynamics are *exogenous*: the subject's ability changes through cognitive or computational processes (learning from feedback, forgetting, continued training), not through strategic optimization of the evaluation instrument itself. This assumption holds in the settings that are the focus of this paper—a student learning from tutoring, a model being fine-tuned on new data, an evaluator tracking ability over time. When measurement carries high-stakes consequences, subjects or their developers may optimize against the instrument rather than developing the underlying capability (Challenge 2); the framework developed here provides the measurement baseline against which such validity threats can be diagnosed using the psychometric tools discussed in Section 7.

The unidimensional simplification. Throughout this paper, we model ability as a scalar $\theta \in \mathbb{R}$. This is a deliberate simplification: in practice, both human abilities and AI capabilities are irreducibly multidimensional. A student may improve in algebra while stagnating in geometry; a language model may gain in code generation while degrading in open-ended reasoning after fine-tuning. Multidimensional IRT (MIRT) extends the framework to vector-valued ability $\boldsymbol{\theta} \in \mathbb{R}^d$, where each item loads on one or more dimensions via a discrimination vector [Reckase, 2009]. This generalization introduces challenges that are absent in the unidimensional case: the geometry of the latent space must be identified (how many dimensions, and what they represent), items must be calibrated with respect to multiple dimensions simultaneously, and adaptive item selection must balance information across dimensions rather than along a single scale [Bonifay, 2019]. These challenges compound in the dynamic regime: when ability is both multidimensional and time-varying, the evaluator must track a trajectory through $\mathbb{R}^d$ rather than along $\mathbb{R}$, and different dimensions may evolve at different rates or even in opposite directions. We adopt the unidimensional formulation because it makes the core structure of each regime transparent—the static/dynamic distinction, the partition between $\mathcal{L}$ and $\mathcal{S}$, and the identifiability arguments all carry over directly to the multidimensional case. We flag multidimensional extensions as open problems within the relevant regimes (Section 5).

Beyond binary item responses. The framework above assumes a binary outcome $y_t \in \{0, 1\}$ for each item, which fits naturally with multiple-choice benchmarks and standardized tests. Modern AI evaluation, however, increasingly relies on response formats that depart from this classical setting. We discuss three important extensions and their implications for the framework.

Pairwise preference data. Systems such as Chatbot Arena [Zheng et al., 2024] evaluate models through pairwise comparisons: two models respond to the same prompt, and a human judge selects the better response. The Bradley–Terry model [Bradley and Terry, 1952] provides the natural link to IRT:

$P(\text{model } i \succ \text{model } j \mid \theta_i, \theta_j) = \sigma(\theta_i - \theta_j)$, which is formally equivalent to the Rasch model with the opponent's ability replacing item difficulty. This means that the core IRT machinery—maximum likelihood estimation, uncertainty quantification, adaptive item selection—transfers directly to pairwise settings, and the Elo rating system is an online approximation to Bradley–Terry maximum likelihood [Boubdir et al., 2023]. The framework applies accordingly: static pairwise tournaments are static measurement; tracking model rankings over time as new versions are released is dynamic measurement; selecting which matchups to present to maximize the discriminative power of the tournament is adaptive testing within either regime. However, pairwise comparisons introduce additional measurement facets—judge identity, prompt selection, and presentation order—each of which contributes variance that a generalizability study should decompose, just as in Challenge 1.

Graded and open-ended responses. Many AI evaluation tasks produce graded or open-ended responses rather than binary outcomes: code generation may be scored by the fraction of test cases passed, translation quality may receive a Likert rating, and open-ended questions may be evaluated by a rubric. The binary assumption can be relaxed via polytomous IRT models such as the graded response model [Samejima, 1969], which models the probability of achieving each score level as a function of ability and item parameters. A more fundamental departure arises when scoring is performed by an LLM-as-judge [Zheng et al., 2024]: the scorer's reliability becomes an additional measurement facet, raising concerns about inter-rater agreement, systematic bias (e.g., preference for longer or more verbose responses), and calibration drift as the judge model itself is updated [Shankar et al., 2024]. These concerns parallel longstanding challenges in human essay scoring, where psychometrics has developed tools for modeling rater effects within generalizability theory—tools that could be imported directly into LLM-as-judge evaluation. A related challenge is that the effective item bank may be unbounded: when items are generated on demand (e.g., by sampling prompts from a distribution), classical assumptions about a fixed, pre-calibrated bank no longer hold, and item parameters must be estimated online or predicted from item features via explanatory IRT [De Boeck and Wilson, 2004]—connecting to the automatic task construction problem discussed in Section 5.

Multi-turn and agentic evaluation. Evaluation paradigms such as SWE-bench [Jimenez et al., 2024], where models must resolve real-world software issues through multi-step interaction with a codebase, challenge the notion of a discrete "item" with a well-defined response. When a task involves planning, tool use, and iterative refinement, the boundary between a single item-response exchange and a sequence of exchanges becomes ambiguous. The framework applies most directly when each multi-step task has a clear terminal success criterion (the issue is resolved or not), so that the task functions as a single binary item despite its internal complexity. For more open-ended interactive settings—such as evaluating conversational quality over a multi-turn dialogue—additional structure is needed to define what constitutes an item and what constitutes a response, and the within-session dynamics ($\mathcal{L}$) may themselves become the object of measurement rather than a confound to be controlled.

We note that some forms of AI evaluation fall outside the scope of this framework entirely. Safety evaluation, red-teaming, and alignment testing involve normative judgments where the "correct" response depends on context, values, and policy—a qualitatively different measurement problem from ability estimation, where items have objectively correct answers or well-defined scoring functions. We flag this boundary explicitly: while the latent-trait machinery could in principle be applied to score distributions over safety-relevant prompts, the construct validity questions are fundamentally different and would require separate treatment.

5 Regimes of Measurement

The two regimes form a natural *progression* of increasing complexity. The simpler case—static measurement—assumes a fixed ability observed in a single session; the inference target is a scalar θ_i. Moving to dynamic measurement introduces ability change: between-session change, within-session change, or both, requiring the evaluator to model a trajectory $\theta_i(t)$ rather than a point. For each regime, we state the formal problem, illustrate it with concrete instantiations from both human assessment and AI evaluation, and identify open directions.

5.1 Static Measurement

In the simplest setting, the subject's ability θ is fixed throughout the evaluation, items are drawn from a calibrated bank, and no feedback is provided. The evaluator's task reduces to classical measurement: estimate θ with minimal measurement effort.

Overview. In human assessment, a plethora of state assessments (such as the California Assessment of Student Performance and Progress), national assessments (like NAEP), and international assessments (such as PISA) rely on IRT to estimate student abilities and evaluate educational outcomes efficiently [Scalise and Gifford, 2006], using calibrated item banks to maximize information about student abilities while minimizing the number of items needed. The most common form of AI evaluation similarly falls under static measurement: a model is evaluated once on a fixed benchmark (e.g., MMLU, HELM, GPQA), producing accuracy scores that serve as point estimates of ability. Martinez-Plumed et al. [2019] apply a 3-parameter logistic IRT to estimate and compare the abilities of 128 machine-learning classifiers, and Polo et al. [2024] show that IRT-based item selection can reduce benchmark size by an order of magnitude while preserving ranking accuracy. However, current AI evaluation practices still largely rely on aggregate scores, which have been criticized for limited ability to predict real-world performance [Bowman and Das, 2021, Wei, 2023]. A deeper problem is *construct validity*: many AI benchmarks claim to measure broad capabilities like "reasoning," but these constructs are often ill-defined and may not correspond to coherent latent traits [Raji et al., 2021]. Additionally, the use of static datasets raises the possibility that evaluation items appear in training data, compromising measurement validity [Ganguli et al., 2023]—motivating the automatic task construction problem (see Open Problem 1 below). The growing emphasis on process-based evaluation—scoring intermediate reasoning steps rather than only final answers [Lightman et al., 2024]—and on harder frontier benchmarks designed to resist saturation [Glazer et al., 2024] both reflect dissatisfaction with current static evaluation, and both connect to the IRT framework: process-based scoring is a form of polytomous or partial-credit IRT (see Section 4), while the construction of harder benchmarks is precisely the item calibration problem of targeting high-difficulty regions of the latent trait scale.

Cross-pollination: from psychometrics to AI evaluation. Two concrete lines of work illustrate how tools developed for human assessment can be imported wholesale into AI evaluation.

Adaptive testing for efficient AI evaluation. Computerized adaptive testing (CAT)—the workhorse of large-scale human assessment for decades—selects items sequentially based on the examinee's estimated ability, achieving the same measurement precision as a full-length test with far fewer items. Truong et al. [2025b] bring this methodology to language model evaluation. They first train a difficulty-prediction model that maps each benchmark item to an IRT difficulty parameter, removing the need for a costly pilot study with real model responses. Given a new model, the system initializes a Rasch IRT model with these predicted difficulties, then runs a standard CAT loop: at each step it selects the item whose difficulty is closest to the current ability estimate, observes the model's response, and updates the ability estimate via maximum likelihood. Across 22 benchmarks and 172 language models, the adaptive procedure recovers benchmark-level accuracy scores within 2% error using only 1–18% of items, and predicts the full ranking of all 172 models (Kendall $\tau \approx 0.96$) from roughly one-tenth of the evaluation budget. The result demonstrates that the same IRT + CAT pipeline used to administer the GRE to millions of human examinees transfers directly to AI evaluation, with the key adaptation being the use of a learned difficulty predictor in place of empirical item calibration.

Psychometric item analysis for benchmark quality. A second import from human testing addresses item quality. In psychometric practice, every operational test undergoes item analysis—statistical diagnostics that flag items behaving anomalously (e.g., items that fail to discriminate between high- and low-ability examinees, or items whose difficulty is misaligned with the target population). Truong et al. [2025a] apply these diagnostics to AI benchmarks. They fit IRT models to the response matrices of nine widely used benchmarks (including MMLU, HellaSwag, and ARC), then compute item-level statistics: tetrachoric correlations between item pairs (to detect redundancy or inconsistency), item-total correlations (to flag items that do not discriminate), and Mokken scalability coefficients (to identify items that violate unidimensionality). Items flagged by these diagnostics are reviewed by human annotators, who confirm that flagged items contain concrete quality problems—ambiguous wording, incorrect answer keys, grading errors, or questions that are unanswerable from the provided context—with up to 84% precision across benchmarks. The study shows that routine psychometric item analysis, applied unchanged from its original human-testing context, can systematically identify quality problems in AI benchmarks that aggregate accuracy scores conceal.

Open Problems

1. **Automatic task construction with calibrated difficulty.** Classical IRT assumes a fixed, pre-calibrated item bank, but modern applications increasingly require constructing new tasks on

demand—for example, to prevent item exposure or to evaluate AI systems on novel problems. Embretson's *cognitive design system* approach [Embretson, 1998] provides a principled foundation: by identifying the cognitive features of a task that determine its difficulty (e.g., the number of reasoning steps, the degree of abstraction), one can predict item parameters from task structure rather than requiring empirical calibration—and the explanatory IRT framework [De Boeck and Wilson, 2004] provides the statistical machinery for embedding such features as item covariates. Recent work on variational IRT [Wu et al., 2020] and generative item modeling [Zelikman et al., 2023] shows that machine learning can produce tasks with approximately controlled difficulty (Fig. 6), but the gap between target and realized difficulty remains substantial. How can we guarantee that automatically constructed tasks are well-calibrated without requiring a full pilot study?

2. **Adaptive sampling for efficient evaluation.** IRT-based adaptive item selection can substantially reduce the number of items needed to achieve a given accuracy in ability estimation compared to random sampling (Fig. 7), making it feasible to assess subjects across multiple benchmarks. However, current adaptive algorithms assume the item pool is well-calibrated and the response model is correctly specified—both assumptions that may fail when evaluating subjects with unfamiliar error patterns. Developing robust adaptive selection methods that account for model misspecification is an open challenge.

3. **Construct validity beyond aggregate accuracy.** Most evaluation protocols report a single aggregate score, implicitly treating all items as measuring the same construct. IRT provides tools for detecting multidimensionality (e.g., items that cluster into separate factors) and item misfit (e.g., items that do not discriminate between high- and low-ability subjects). When multidimensionality is detected, MIRT [Reckase, 2009, Bonifay, 2019] can model it explicitly, estimating how each item loads on multiple latent dimensions and producing dimension-specific ability profiles rather than a single score. Systematically applying these diagnostics could reveal which evaluation instruments measure coherent constructs and which conflate distinct capabilities—and, where benchmarks are multidimensional, whether reporting a single aggregate score masks meaningful variation across dimensions. Truong et al. [2025a] provide a concrete demonstration: applying item-level diagnostics (tetrachoric correlations, item-total correlations, Mokken scalability) to nine AI benchmarks, they flag invalid items—ambiguous wording, incorrect answer keys, grading errors—with up to 84% precision, showing that routine psychometric item analysis can identify quality problems that aggregate scores conceal.

4. **Item calibration.** Throughout this paper we assume the item bank is calibrated—item parameters z_t are known to the evaluator—which is standard in psychometric practice. However, AI benchmarks are rarely calibrated in the IRT sense: items are typically selected by subject-matter experts or crowdsourced without a formal piloting process. Extending the framework to handle joint calibration (estimating item and ability parameters simultaneously) or online calibration (updating item parameters as new response data arrives) is an important open direction [Baker, 2001, Wainer and Mislevy, 2000].

5.2 Dynamic Measurement

Once ability is no longer static, the evaluator must model how it changes over time. Two categories of change, partitioned along the session boundary, may be at play: *within-session change* via $\theta_{t+1} = \mathcal{L}_\tau(\theta_t, z_t, y_t)$, where the evaluation process itself alters the subject's ability, and *between-session change* via $\theta_{t+s} = \mathcal{S}_\psi(\theta_t)$, where unobserved processes alter ability during gaps in evaluation. These can appear alone or together. The evaluator's goal is to estimate the trajectory $\theta(t)$ and infer the parameters governing the dynamics (τ, ψ, or both).

The relative prominence of between-session and within-session change depends on the application. In some cases only one is present—longitudinal tracking without feedback involves only between-session change, while a single feedback-rich session involves only within-session change. In many practical settings both are at play, and disentangling them becomes a central challenge.

Between-session drift. When measurements are taken across sessions with gaps between them, ability may change for reasons unobserved by the evaluator. The primary objective is to understand how ability evolves over time, often referred to as the "growth curve" [Tripathi and Domingue, 2019]. There are multiple realizations of this sub-case, reflecting different data collection scenarios:

1. **Regular time points with sufficient data.** In some studies, N assessments are administered at N distinct time points, with enough items at each point to obtain a reliable cross-sectional estimate of ability [Embretson, 1991]. The trajectory is then modeled by fitting $\mathcal{S}_\psi$ to the sequence of point estimates. This is the most data-rich case and permits standard growth-curve methods.

2. **Sparse or irregular data collection.** In other scenarios, data collection may occur sporadically—for example, a handful of items administered at irregular intervals—such that no single time point provides a reliable estimate of ability. This setting requires models that jointly estimate the trajectory and individual time-point abilities, borrowing strength across time via the smoothness of between-session change.

3. **Pre–post measurement.** A common design involves collecting items at just two time points: before and after an intervention. The focus is on estimating the change in ability, $\Delta\theta = \theta_{\text{post}} - \theta_{\text{pre}}$. While this design identifies the net effect, it provides no information about the trajectory between the two points—ability could have changed monotonically, oscillated, or remained flat until a sudden shift.

Across all three scenarios, the core inference problem is the same: recovering a continuous ability trajectory from discrete, noisy observations. Many latent dynamics models can serve this role—state-space models, Gaussian processes (GPs), neural ODEs—each combining an observation model (e.g., IRT likelihood) with a prior over smooth trajectories, and differing in flexibility, computational cost, and interpretability. In Appendix B, we illustrate one concrete approach using a GP framework and present a simulation study showing that distributing a fixed item budget across more time points yields finer temporal resolution of the ability trajectory (Fig. 8), at the cost of less precise estimates at each individual time point. We also describe an alternative, computationally simpler approach based on the Elo rating system [Pelánek, 2016].

Within-session learning. When feedback is present, learning can occur within a session. One productive framing treats learning as updating an internal model: the subject continuously integrates prior knowledge with new evidence to revise its beliefs. This Bayesian perspective has been influential in cognitive science [Anderson, 1990, Tenenbaum et al., 2006, Griffiths and Tenenbaum, 2005] and is increasingly relevant for understanding how AI systems adapt during evaluation (e.g., in-context learning). Under this view, the subject's ability changes in response to feedback via some learning dynamic $\theta_{t+1} = \mathcal{L}_\tau(\theta_t, z_t, y_t)$. Several classes of learning rules are natural candidates:

- **Proportional feedback update (gradient ascent):** $\theta_{t+1} = \theta_t + \tau \frac{\partial L}{\partial \theta_t}$, where $L = y_t \log P(y_t = 1|\theta, z_t) + (1 - y_t) \log(1 - P(y_t = 1|\theta, z_t))$ is the log-likelihood and τ is a learning rate. This is the simplest parametric choice: the subject adjusts ability in the direction that makes the observed outcome more likely, with a fixed step size.

- **Evidence accumulation (Bayesian updating):** The subject maintains a posterior distribution $p(\theta_t|y_{1:t})$ and updates it upon receiving feedback, with the point estimate given by the posterior mean or MAP. This yields a learning rule where the effective step size shrinks as more evidence accumulates, naturally capturing diminishing returns from practice.

- **History-sensitive update (momentum):** $\theta_{t+1} = \theta_t + \tau \frac{\partial L}{\partial \theta_t} + \beta(\theta_t - \theta_{t-1})$, where the momentum term β allows learning to depend on the recent trajectory, not just the current feedback. This can capture phenomena like "learning streaks," where consecutive successes accelerate improvement.

The choice of learning rule is itself a modeling decision with empirical consequences: different rules predict different responses to the same sequence of items and feedback, and misspecifying the rule can bias both ability estimates and item selection.

Overview. In human assessment, dynamic measurement arises across diverse contexts. Vandemeulebroecke et al. [2017] apply drift-only measurement to neuropsychiatric monitoring, administering tests every six months and modeling between-session ability change as a linear function of subject-specific covariates. Martin and Quinn [2002] measure the ideal points of Supreme Court justices over time using a Gaussian random walk. In settings with feedback, Abbakumov et al. [2019] propose dynamic extensions of the Rasch model for MOOCs, decomposing ability change into continuous growth from lectures ($\mathcal{S}$) and local growth from repeated attempts ($\mathcal{L}$), while Klinkenberg et al. [2011], Kadengye et al. [2014],

and Bolsinova et al. [2022] study students on adaptive learning platforms, illustrating a progression from Elo-based methods through models that include session-level terms for between-session changes—though none fully disentangle between-session from within-session change.

In AI evaluation, dynamic measurement manifests in three forms. Foundation models are periodically updated (e.g., GPT-3.5 → GPT-4 → GPT-4o), creating between-session ability change at discrete checkpoints—though AI capability trajectories may be less smooth than those typically assumed in human psychometrics, as a major architecture change can produce discontinuous jumps. In-context learning creates within-session change: when a model receives examples within its context window, its effective ability changes as a function of the prompts it has seen, manifesting as sensitivity to prompt ordering and few-shot example selection [Lu et al., 2022]. Position effects—where the same item becomes easier or harder depending on where it appears in the evaluation sequence—have been formally modeled in human assessment through position-sensitive IRT [Kanopka and Domingue, 2025], which disentangles fatigue and practice effects from true ability; analogous prompt-ordering effects in AI evaluation await similar treatment. Models fine-tuned between periodic evaluations combine both sources: between-session change from continued training and within-session change from task exposure, with benchmark contamination as a particularly acute instance.

Cross-pollination: ML-native trajectory models for human measurement. The mirror of the transfer described under Static Measurement above—psychometric tools imported into AI evaluation—is the transfer of AI's expressive trajectory-modeling tools into human dynamic measurement. Classical dynamic psychometrics has relied on parsimonious families: linear growth, random-walk drift, Rasch-with-time. Machine learning brings a more expressive toolkit—Gaussian processes with learned kernels, state-space models with neural latents, attention-based sequence models—that can capture trajectory structure the simpler families cannot express. Truong et al. [2025c] provide one concrete instantiation. They collect a dataset of 3,286 undergraduate students solving 396 C++ programming problems across six weeks of lab sections, recording over 3 million submissions with test-case-level binary outcomes; the median student submits 455 times with a median inter-submission interval of 1.1 minutes, producing rich within-session learning trajectories. To these trajectories they fit five dynamic IRT models spanning the classical-to-ML-native progression: (1) Elo, which absorbs all dynamics into a single update rule; (2) a Rasch model with linear growth, $\theta_{st} = \theta_{0s} + \gamma_s \cdot t + \sigma \cdot \eta_{st}$; (3) Gaussian Process IRT (GPIRT), which places a GP prior over the ability trajectory with a kernel encoding time between submissions; (4) a Changepoint IRT (CIRT) model; and (5) a Recurrent State Space Model (RSSM) that uses a GRU to encode the full submission history into a latent state. The study demonstrates that the ML-native families (GPIRT, CIRT, RSSM) apply directly to human response data and sit in the same model-comparison pipeline alongside classical dynamic IRT, validating the transfer direction for trajectories—like programming submissions at minute-scale intervals—where classical families would be too coarse. The same infrastructure also surfaces a natural follow-up: whether LLMs prompted to simulate students can generate synthetic trajectories that preserve these statistical properties. Truong et al. [2025c] report a negative answer for the programming case—models trained predominantly on correct code reproduce ability structure only weakly (best $\rho \approx 0.36$) and fail to match the novice error patterns that characterize real learning—a tension between functional correctness and behavioral fidelity that limits synthetic-data approaches to dynamic calibration.

Open Problems

1. **Model specification for dynamics.** Different applications use different models for between-session change (linear, random walk, GP with a specific kernel) and within-session change (proportional, Bayesian, history-sensitive). How should the practitioner choose? Bayesian model comparison (e.g., via marginal likelihoods) could provide a principled answer, but has not been systematically applied. Relatedly, can the within-session dynamics themselves be identified from response data—and does the answer depend on the richness of the item sequence and the length of the session?

2. **Disentangling between-session from within-session change.** As discussed in Section 4, when both $\mathcal{L}$ and $\mathcal{S}$ are present, the two processes may not be separately identifiable from response data alone. Existing approaches either absorb both dynamics into a single update or model one while ignoring the other. The conditions outlined in Section 4—no-feedback measurement probes, session boundary variation, and parametric smoothness assumptions—provide starting points, but

a formal identifiability analysis (e.g., establishing necessary and sufficient conditions under specific model classes) remains open.

3. **Optimal measurement schedules.** Given a budget of N total items, how should they be distributed across time to minimize uncertainty about the ability trajectory? The simulation in Fig. 8 provides initial evidence, but a formal treatment connecting to experimental design theory (e.g., D-optimal designs for GP hyperparameters) is lacking. The benefit of feedback may also depend on when it is delivered relative to breaks, and consolidation during breaks may depend on the type of feedback received [Cepeda et al., 2009].

4. **Non-stationary dynamics.** The discussion above assumes between-session and within-session change are stationary (the same process governs ability changes at all times). In practice, regime changes occur—a student begins tutoring, a patient starts medication, an AI model receives a major architecture update. Detecting such changepoints from response data alone remains an open challenge.

5. **Feedback modality.** Most models assume binary feedback (correct/incorrect), but real settings involve graded feedback, hints, worked examples, and scaffolding. Each modality implies a different within-session dynamic: a hint may shift ability less than a full explanation, and repeated exposure to worked examples may follow a different learning curve than trial-and-error with binary feedback. Extending the framework to handle heterogeneous feedback types is largely unexplored.

6. **In-context learning as dynamic measurement.** When an AI model receives few-shot examples before evaluation, the examples function as feedback that changes the model's effective ability within the evaluation session. This is *formally* analogous to within-session learning—both can be modeled as $\theta_{t+1} = \mathcal{L}_\tau(\theta_t, z_t, y_t)$—but the underlying mechanisms differ in important ways: a student learning from feedback updates long-term memory representations, while a language model performing in-context learning conducts inference over a fixed parameter set with an extended context window. The "learning" is entirely determined by the attention mechanism and is not retained beyond the context window. These mechanistic differences may affect which measurement models are appropriate: for instance, in-context learning may exhibit sharper saturation effects and lack the consolidation dynamics that characterize human learning. Developing IRT-based methods that account for in-context learning dynamics [Ross and Andreas, 2024] could provide principled tools for disentangling baseline ability from the boost provided by the prompt, but such methods should be validated against the specific phenomenology of in-context learning rather than assumed to transfer directly from human within-session models.

7. **Longitudinal item calibration.** When a subject encounters the same item across multiple sessions, the item's effective difficulty may change (due to memory, familiarity, or strategy development). Standard IRT assumes fixed item parameters, but in dynamic settings, item parameters may need to co-evolve with subject ability—a challenge that connects to the task construction problem raised under static measurement.

6 Coupled Human–AI Measurement

The parallel instantiations in Section 5 show that human assessment and AI evaluation face *formally analogous* problems: both involve inferring a latent ability from item responses, and both confront the same core challenges when that ability changes over time. The parallels are tightest where the formal structure genuinely matches—benchmark contamination and teaching to the test share the same mechanism (exposure to evaluation content inflating measured performance)—and looser where the mechanisms diverge (in-context learning involves transient inference over a fixed model, whereas human within-session learning involves durable memory updates). Despite these mechanistic differences, the *measurement problems* that arise are structurally similar, and solutions developed in one domain can serve as starting hypotheses for the other—though validating whether a solution actually transfers requires empirical work in the target domain.

In many emerging applications, however, these are not separate problems at all—human and AI subjects interact in the same system, and their measurement problems become formally coupled. Some of these systems are themselves measurement pipelines; others are interactional in character but surface a measurement question as a subproblem. In every case our focus is *inference*, not intervention: when

both subjects' trajectories shift and the observation process for one depends on the state of the other, neither trajectory can be identified without a joint model. The scenarios below sketch a *research agenda* for this coupled-measurement problem; to our knowledge, no existing work has attempted to jointly recover the trajectories of coupled human and AI subjects within a unified framework.

LLM-as-judge. An increasingly common scoring pipeline replaces automatic graders with an LLM judge: Chatbot Arena [Zheng et al., 2024] uses human judges at scale, and many benchmarks now use GPT-4 or Claude as the judge for open-ended outputs (code quality, summarization, reasoning chains, dialogue). The measurement problem is coupled by construction. The subject model has an unknown capability trajectory $\theta_{\text{subj}}(t)$; the judge has its own latent properties—reliability, calibration, known biases such as preference for longer outputs or self-preference when judging outputs from its own family—and these properties themselves drift as the judge model is updated or as human annotators accumulate exposure [Shankar et al., 2024]. What the pipeline records is the judge's score of the subject's output, which depends on both states. A change in scores between two time points can reflect improvement in the subject, drift in the judge, or any combination; the two are not separately identifiable without a joint model. Classical psychometrics has partial tools: generalizability theory treats raters as a variance-decomposable facet, and rater-effect models in IRT estimate rater severity and reliability alongside examinee ability. These tools are, to our knowledge, not systematically applied to LLM-as-judge evaluation, despite that pipeline now operating at massive scale with no underlying theoretical framework. This is the purest instance of coupled measurement: both subjects drift through independent processes, no one is being deliberately trained through the measurement, and yet valid inference requires modeling the coupling.

AI tutoring humans. In AI-tutored classrooms—already widespread and growing rapidly—the human learner and the AI tutor are *both* dynamic subjects in the same system. The student's ability evolves through interaction with the tutor (within-session change) and through unobserved processes between sessions (between-session change), while the tutor's own behavior changes as it adapts to student responses, accumulates interaction data, or receives model updates. The measurement question is whether observed changes in student performance reflect genuine learning, the tutor adaptively easing items, or the tutor's own capability shifting—and these are confounded without a joint model of both trajectories. Evaluating the tutor's effectiveness requires the same joint model, run in the other direction. The intervention context (tutoring) is incidental to the framework; the formal structure is the same coupled-inference problem that appears in the LLM-as-judge case, with the student playing the role of subject and the tutor playing the role of measurement instrument.

Humans training AI. The coupling runs in the opposite direction in reinforcement learning from human feedback (RLHF), red-teaming, and human-in-the-loop curriculum design: the human serves as evaluator while the AI model is the subject. Again the paper's concern is not the training itself but the fact that any measurement extracted from the interaction inherits the coupling. The human annotator's judgment is not static—internal criteria sharpen through exposure, expectations calibrate to the model's evolving capability, and fatigue or anchoring introduce between-session drift in annotation quality—while the model's capability changes in response to the very judgments being provided. Reward-model reliability, preference-label validity, and downstream capability claims are all quantities derived from this coupled observation process; none is identifiable without a joint model of the annotator's trajectory and the model's. This is the mirror image of the AI-tutored classroom, and the same formal tools—jointly modeling between-session and within-session change for both subjects—would apply.

7 Looking Forward

This roadmap has proposed a framework organized around two predictive-measurement regimes—static measurement, inferring a scalar ability θ_i, and dynamic measurement, inferring a trajectory $\theta_i(t)$—that structure the landscape of measurement for ability. The parallel instantiations from human assessment and AI evaluation within each regime reveal that the two fields face formally analogous challenges—sharing the same abstract structure of inferring latent ability from item responses, even where the underlying mechanisms diverge (see the discussion of tight vs. loose parallels in Section 5). As Figure 1 illustrates, the two fields bring complementary strengths that can address each other's limitations. We close by discussing these cross-pollination opportunities and the shared frontier that unites them.

From psychometrics to AI evaluation (Figure 1a, top arrow). Psychometrics brings three foundational strengths that directly address the most pressing challenges in AI evaluation. First, *construct theory*—the systematic practice of defining what is being measured before measuring it—provides a principled approach to a problem that increasingly plagues AI benchmarking: many benchmarks claim to measure broad capabilities like "reasoning" or "common sense," yet these constructs are often ill-defined and may not correspond to coherent latent traits. Psychometric tools such as factor analysis and item-construct alignment can reveal whether a benchmark measures a single coherent ability or conflates distinct capabilities. Second, *validity and invariance frameworks* offer established methodologies for determining whether an evaluation actually measures what it claims to measure and whether scores are comparable across different populations and conditions—crucial for AI evaluation, where the same benchmark is applied to models with vastly different architectures and training regimes. Generalizability theory [Cronbach et al., 1963] is a particularly actionable example: a G-study applied to AI benchmarking would treat models as subjects and cross them with facets such as prompt format (zero-shot, few-shot, chain-of-thought), item ordering, temperature setting, and scoring rubric, decomposing variance to quantify how much of benchmark score variation reflects genuine model differences versus incidental facets of the evaluation protocol. To our knowledge, no such G-study has been conducted for AI benchmarks despite the data being readily available. Third, *principled test design*—including item calibration, adaptive testing, and test equating—can transform current AI benchmarking practices, which rely heavily on aggregate accuracy scores from fixed item sets, into rigorous measurement instruments. Truong et al. [2025a] provide a concrete demonstration of this transfer in action: applying standard psychometric item diagnostics—tetrachoric correlations, item-total correlations, and Mokken scalability analysis, all grounded in classical test theory and the Rasch model—to nine widely used AI benchmarks (including GSM8K, MMLU, and MedQA), they identify invalid items with up to 84% precision. The flagged problems—ambiguous wording, incorrect answer keys, and grading errors—are invisible to aggregate accuracy scores but distort model rankings (e.g., shifting a model from near-bottom to near-top of the leaderboard after correction). This work illustrates the kind of cross-pollination this roadmap advocates: routine psychometric tools, applied without modification, reveal quality problems in AI benchmarks that the AI evaluation community had not systematically detected.

From AI to human assessment (Figure 1a, bottom arrow). Conversely, the AI evaluation ecosystem brings strengths that address longstanding limitations of human psychometrics. First, *powerful predictive models*—including deep neural networks and attention-based architectures—can capture response patterns far more complex than those assumed by classical IRT, potentially revealing structure in human assessment data that parametric models miss. Second, *rich, fast temporal data* from AI training pipelines (model checkpoints, in-context learning trajectories, continual training logs) provide densely sampled ability trajectories that are impossible to obtain in human studies; these trajectories serve as testbeds for developing and validating dynamic measurement methods that can then be transferred to the sparser longitudinal data available in education and clinical assessment. Third, *scalable inference methods*—variational inference, amortized estimation, and parallel computation—make it feasible to fit sophisticated measurement models to datasets orders of magnitude larger than those traditionally used in psychometrics, enabling real-time adaptive assessment at scale. Importantly, these two directions of transfer form a virtuous cycle rather than independent one-way exchanges: psychometric methodology imported into AI evaluation (the top arrow) can itself be executed at unprecedented scale using AI capabilities (the bottom arrow). Construct validity assessment—traditionally a labor-intensive process requiring expert item review, pilot testing, and manual diagnostic analysis—could be conducted across large benchmark suites through human-AI collaboration, combining psychometric diagnostics with LLM-assisted item analysis to enable systematic quality assurance that neither field could achieve alone.

The shared frontier: tracking ability over time (Figure 1b). The temporal dimension represents the most promising frontier for collaboration. Neither field has fully solved the problem of tracking ability as it evolves over time—in response to interventions, feedback, and the measurement process itself. In human assessment, longitudinal data are sparse, growth-curve methods are often limited to simple parametric forms, and the interaction between learning and forgetting remains poorly understood. In AI evaluation, the challenge is different but structurally analogous: model capabilities shift rapidly across versions and training stages, evaluations are typically one-shot snapshots that miss the trajectory, and the entanglement of evaluation with training—through benchmark contamination or in-context learning—mirrors the confounding of measurement with learning in human settings. The Gaussian process framework developed in this paper (Appendix B) provides one approach to this shared problem,

but much remains open: detecting regime changes from response data alone, jointly identifying learning and drift processes, designing evaluation protocols that are robust to the dynamic nature of the subjects they measure, and extending the assumption of a calibrated item bank—standard in psychometrics but rarely satisfied in AI evaluation—to support online or joint calibration. Progress on any of these fronts—whether initiated in the human or AI domain—benefits both. Both fields also share an underlying response-matrix structure that connects them to collaborative filtering and matrix completion, yet neither has fully exploited this connection for the dynamic settings where measurement and learning interact.

Concrete next steps. A roadmap should identify not only what problems exist but how to make progress on them. We highlight two near-term research directions—one per measurement regime—that would provide immediate empirical grounding.

1. *Static measurement.* Recent work has begun fitting IRT models to AI benchmark data—Zhou et al. [2025] apply a 4PL model to 11 benchmarks, and Polo et al. [2024] use IRT for efficient benchmark reduction—and factor-analytic studies have examined benchmark structure at the aggregate level [Ilic and Gignac, 2024, Burnell et al., 2023]. However, these efforts have largely borrowed the estimation machinery of IRT while skipping the assumption-checking diagnostics that are standard in psychometric practice. Truong et al. [2025a] take a significant step in this direction, showing that item-level diagnostics rooted in classical test theory and the Rasch model—tetrachoric correlations, item-total correlations, and Mokken scalability—can identify invalid benchmark items with up to 84% precision across nine benchmarks, revealing problems (ambiguous wording, incorrect keys, grading errors) that distort model rankings and are invisible to aggregate scores. The natural next step is to extend this item-level scrutiny to the full suite of psychometric assumption-testing diagnostics: local independence (do item responses exhibit dependencies beyond what the latent trait explains?—no study has reported Q3 statistics or residual correlations), item-level dimensionality within individual benchmarks (existing factor analyses operate on benchmark-aggregate scores, not on items within a single benchmark), and differential item functioning across model families (do items function differently for, say, open-weight vs. proprietary models after controlling for overall ability?—entirely untested, though evidence that model embeddings cluster by architecture family suggests DIF would be detected). This diagnostic work requires only publicly available response data and standard psychometric software, making it immediately feasible.
2. *Dynamic measurement.* Fitting the GP-IRT framework developed in Appendix B—or a comparable latent dynamics model—to longitudinal AI evaluation data (e.g., model checkpoint evaluations during training, or Chatbot Arena ratings over time) would test whether smooth latent trajectories can be recovered from noisy benchmark observations and whether the inferred trajectories reveal structure not visible in raw scores, such as phase transitions, plateaus, or post-training regressions. An analogous study using longitudinal student data from an adaptive learning platform (e.g., ASSISTments or Duolingo) would test the same methods in the human domain.

Each of these studies is feasible with existing data and methods; together, they would transform the framework from an organizing scheme into an empirically grounded research program.

One important direction this roadmap does not address is the strategic dimension: when measurement carries high-stakes consequences, subjects or their developers may optimize against the evaluation instrument rather than developing the underlying capability [Goodhart, 1984]. The framework developed here assumes exogenous dynamics (Section 4) and provides the measurement baseline against which such validity threats can be detected, but does not model the strategic interaction itself.

Without this unified measurement science, education and AI development will continue to rely on evaluation signals that do not reliably distinguish genuine ability from artifacts of the testing process. We hope that this roadmap—and the explicit identification of cross-pollination opportunities in Figure 1—will encourage collaboration between the psychometrics and machine learning communities and accelerate progress on the fundamental problem of measuring and improving intelligent systems.

References

Dmitry Abbakumov, Pieter Desmet, and Wim Van den Noortgate. Measuring growth in students' proficiency in moocs: Two component dynamic extensions for the rasch model. *Behavior Research Methods*, 51(1):332–341, February 2019. doi: 10.3758/s13428-018-1129-1.

Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11), 2023. ISSN 0360-0300. doi: 10.1145/3569576.

John R. Anderson. *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, 1990.

Heidi L Andrade and Gregory J Cizek. *Handbook of Formative Assessment*. Routledge, 2010.

Yigal Attali. Immediate feedback and opportunity to revise answers: Application of a graded response irt model. *Applied Psychological Measurement*, 35(6):472–479, 2011. ISSN 0146-6216. doi: 10.1177/0146621610381755.

Frank B Baker. *The basics of item response theory*. ERIC, 2001.

Andrew M. Bean, Luc Rocher, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, 2025.

James Bennett and Stan Lanning. The Netflix prize. *Proceedings of KDD Cup and Workshop*, 2007.

Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74, 1998.

Maria Bolsinova, Matthieu JS Brinkhuis, Abe D Hofman, and Gunter Maris. Tracking a multitude of abilities as they develop. *British Journal of Mathematical and Statistical Psychology*, 75(3):753–778, 2022.

Wes Bonifay. *Multidimensional item response theory*. Sage Publications, 2019.

Meriem Boubdir, Edward Bouchard, Beyza Kamber, Sanmi Koyejo, and Alex Dimakis. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*, 2023.

Samuel R. Bowman and George Das. What will it take to fix benchmarking in natural language understanding?, 2021. URL https://arxiv.org/abs/2104.02145.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138, 2023. doi: 10.1126/science.adf6369. URL https://www.science.org/doi/abs/10.1126/science.adf6369.

Robert J. Carroll, David M. Primo, and Brian K. Richter. Using item response theory to improve measurement in strategic management research: An application to corporate social responsibility. *Strategic Management Journal*, 37(1, SI):66–85, 2016. ISSN 0143-2095. doi: 10.1002/smj.2463.

Devin Caughey and Christopher Warshaw. Dynamic estimation of latent opinion using a hierarchical group-level irt model. *Political Analysis*, 23(2):197–211, 2015. ISSN 1047-1987. doi: 10.1093/pan/mpu021.

Nicholas J Cepeda, Noriko Coburn, Doug Rohrer, John T Wixted, Michael C Mozer, and Harold Pashler. Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental psychology*, 56(4):236–246, 2009.

HH Chang and ZL Ying. ¡i¿a¡/i¿-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3):211–222, 1999. ISSN 0146-6216. doi: 10.1177/01466219922031338.

Youngduck Choi, Youngnam Lee, Junghyun Shin, Jineon Cho, Seoyon Park, Seewoo Lee, Jongwon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the 7th ACM Conference on Learning @ Scale*, pages 341–344, 2020.

Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994.

Lee J Cronbach, Nageswari Rajaratnam, and Goldine C Gleser. Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16(2):137–163, 1963.

Paul De Boeck and Mark Wilson. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York, 2004.

Martijn G. De Jong, Jan-Benedict E. M. Steenkamp, Jean-Paul Fox, and Hans Baumgartner. Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, 45(1):104–115, FEB 2008. ISSN 0022-2437. doi: 10.1509/jmkr.45.1.104.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Benjamin W. Domingue, Klint Kanopka, Logan Caffrey-Maffei, Joshua B. Gilbert, Radhika Kapoor, Yutong Liu, Susie Nadela, Guanyu Pan, Ling Zhang, Shuai Zhang, Mika Braginsky, and Michael C. Frank. An introduction to the item response warehouse (IRW): A resource for enhancing data usage in psychometrics. *Behavior Research Methods*, 57:276, 2025.

F. Y. Edgeworth. The statistics of examinations. *Journal of the Royal Statistical Society*, 51(3):599–635, 1888. ISSN 09528385. URL http://www.jstor.org/stable/2339898.

Susan E Embretson. A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3):495–515, 1991.

Susan E. Embretson. A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3(3):380–396, 1998.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

Deep Ganguli, Nicholas Schiefer, Marina Favaro, and Jack Clark. Challenges in evaluating AI systems, 2023. URL https://www.anthropic.com/index/evaluating-ai-systems.

Aritra Ghosh, Neil Heffernan, and Andrew S Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2330–2339, 2020.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, and Anson Ho. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Charles AE Goodhart. Problems of monetary management: The UK experience. *Monetary Theory and Practice*, pages 91–121, 1984.

Thomas L. Griffiths and Joshua B. Tenenbaum. Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773, 2005.

Ronald K Hambleton and Russell W Jones. Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice*, 12 (3):38–47, 1993.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, CA, 1991.

Jose Hernandez-Orallo. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, 2017.

Paul W. Holland and Howard Wainer. *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

David Ilic and Gilles E. Gignac. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858, 2024.

Shengyu Jiang, Jiaying Xiao, and Chun Wang. On-the-fly parameter estimation based on item response theory in item-based adaptive learning systems. *Behavior Research Methods*, 55(6):3260–3280, 2023.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Damazo T Kadengye, Eva Ceulemans, and Wim Van den Noortgate. A generalized longitudinal mixture irt model for measuring differential growth in learning environments. *Behavior research methods*, 46: 823–840, 2014.

Hyeon-Ah Kang, Adam Sales, and Tiffany A Whittaker. Flow with an intelligent tutor: A latent variable modeling approach to tracking flow during artificial tutoring. *Behavior Research Methods*, 56(2):615–638, 2024.

Klint Kanopka and Benjamin W. Domingue. A position-sensitive mixture item response model. *Journal of Educational and Behavioral Statistics*, 50(1), 2025.

Radhika Kapoor, Sang T. Truong, Nick Haber, Maria Araceli Ruiz-Primo, and Benjamin W. Domingue. Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint arXiv:2502.20663*, 2025.

Sharon Klinkenberg, Marthe Straatemeier, and Han LJ van der Maas. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824, 2011.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

Charles Lang, George Siemens, Alyssa Wise, and Dragan Gašević. *Handbook of Learning Analytics*. Society for Learning Analytics Research, 2017.

D. N. Lawley. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences*, 61(3):273–287, 1943. doi: 10.1017/S0080454100006282.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2024.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8086–8098, 2022.

William A. Ma, Adam Richie-Halford, Amy K. Burkhardt, Klint Kanopka, Clancy Chou, Benjamin W. Domingue, and Jason D. Yeatman. ROAR-CAT: Rapid online assessment of reading ability with computerized adaptive testing. *Behavior Research Methods*, 57:56, 2025.

Andrew D. Martin and Kevin M. Quinn. Dynamic ideal point estimation via markov chain monte carlo for the u.s. supreme court, 1953–1999. *Political Analysis*, 10(2):134–153, 2002. doi: 10.1093/pan/10.2.134.

Fernando Martinez-Plumed, Ricardo B. C. Prudencio, Adolfo Martinez-Uso, and Jose Hernandez-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42, 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.09.004.

Naaz Modan. Montana scores rare federal testing waiver in favor of through-year assessment, 2023. URL https://www.k12dive.com/news/montana-federal-waiver-standardized-summative-assessment-through-year-assessment\protect\discretionary{\char\hyphenchar\font}{}{}accountability/690644/.

Peter CM Molenaar. A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, 50(2):181–202, 1985.

Pilar Munday. The case for using duolingo as part of the language classroom experience. *Digital Commons at SHU*, 2016.

Radek Pelánek. Applications of the elo rating system in adaptive educational systems. *Computers & Education*, 98:169–179, 2016.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Anna N Rafferty, Emma Brunskill, Thomas L Griffiths, and Patrick Shafto. Faster teaching via pomdp planning. *Cognitive Science*, 40(6):1290–1332, 2016.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

Mark D. Reckase. *Multidimensional Item Response Theory*. Statistics for Social and Behavioral Sciences. Springer, New York, NY, 2009. doi: 10.1007/978-0-387-89976-3.

Henry L Roediger and Jeffrey D Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255, 2006.

Alexis Ross and Jacob Andreas. Toward in-context teaching: Adapting examples to students' misconceptions, 2024. URL https://arxiv.org/abs/2405.04495.

André A. Rupp, Jonathan Templin, and Robert A. Henson. *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press, New York, 2010.

Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2):1–97, 1969.

Kathleen Scalise and Bernard Gifford. Computer-based assessment in e-learning: A framework for constructing" intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning and Assessment*, 4(6), 2006.

Michael Scriven. The methodology of evaluation. In Ralph W. Tyler, Robert M. Gagné, and Michael Scriven, editors, *Perspectives of Curriculum Evaluation*, pages 39–83. Rand McNally, 1967.

Shreya Shankar, J.D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. Who validates the validators? aligning LLM-assisted evaluation of LLM outputs with human preferences. *arXiv preprint arXiv:2404.12272*, 2024.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 1904. URL https://psycnet.apa.org/record/1926-00292-001.

Joshua B. Tenenbaum, Thomas L. Griffiths, and Charles Kemp. Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318, 2006.

Ajay Tripathi and Benjamin Domingue. Curve fitting from probabilistic emissions and applications to dynamic item response theory. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1336–1341, 2019. doi: 10.1109/ICDM.2019.00170.

Sang Truong, Yuheng Tu, Michael Hardy, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Ben Domingue, Nick Haber, and Sanmi Koyejo. Fantastic bugs and where to find them in AI benchmarks. *arXiv preprint arXiv:2511.16842*, 2025a.

Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025b.

Sang Truong et al. Measuring learners via latent dynamic models. *In preparation*, 2025c.

Wim J. van der Linden. Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2): 201–216, 1998. ISSN 0033-3123. doi: 10.1007/BF02294775.

Wim J Van der Linden, Cees AW Glas, et al. *Computerized adaptive testing: Theory and practice*, volume 13. Springer, 2000.

Marc Vandemeulebroecke, Bjorn Bornkamp, Tillmann Krahnke, Johanna Mielke, Andreas Monsch, and Peter Quarg. A longitudinal item response theory model to characterize cognition over time in elderly subjects. *CPT:Pharmacometrics & Systems Pharmacology*, 6(9):635–641, 2017. ISSN 2163-8306. doi: 10.1002/psp4.12219.

Matthias Von Davier, Xueli Xu, and Claus H Carstensen. Using the general diagnostic model to measure learning and change in a longitudinal large-scale assessment. *ETS Research Report Series*, 2009(2): i–22, 2009.

Howard Wainer and Robert J Mislevy. Item response theory, item calibration, and proficiency estimation. In *Computerized adaptive testing*, pages 61–100. Routledge, 2000.

Howard Wainer, Neil J Dorans, Ronald Flaugher, Bert F Green, and Robert J Mislevy. *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum Associates, 2nd edition, 2000.

Walter D. Way and Clyde M. Reese. An investigation of the use of simplified irt models for scaling and equating the toefl test. *ETS Research Report Series*, 1990(2):i–22, 1990. doi: https://doi.org/10.1002/j.2333-8504.1990.tb01365.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1990.tb01365.x.

Jason Wei. Evaluations: How can we evaluate ai systems?, 2023. URL https://www.jasonwei.net/blog/evals.

Mike Wu, Richard L. Davis, Benjamin W. Domingue, Chris Piech, and Noah Goodman. Variational item response theory: Fast, accurate, and expressive, 2020. URL https://arxiv.org/abs/2002.00276.

Yanbo Xu and Jack Mostow. Using item response theory to refine knowledge tracing. In *Educational Data Mining 2013*, 2013.

Eric Zelikman, Wanjing Anya Ma, Jasmine E. Tran, Diyi Yang, Jason D. Yeatman, and Nick Haber. Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency, 2023. URL https://arxiv.org/abs/2310.06837.

Haoran Zhang and Feiming Li. Dynamic state space models for item response theory. *Applied Psychological Measurement*, 46(2):140–155, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *arXiv preprint arXiv:2505.15055*, 2025.

Yan Zhuang, Qi Liu, Yuting Ning, Wei Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, and Enhong Chen. From static benchmarks to adaptive testing: Psychometrics in ai evaluation. *arXiv preprint arXiv:2306.10512*, 2023.

Study	Δ_t	N_t	Pure Measurement	Item parameters known
Pelánek [2016]	none	10-100 (p.175, para.3)	No (p.176, para.1)	Yes (p.175, para.6)
Martin and Quinn [2002]	1 year	40-104 (p.137, para.2)	Yes	No
Jiang et al. [2023]	none	20-30 (p.3267, para.4)	Yes	Yes (p.3267, para.7)
Kang et al. [2024]	variable	∼50 (p.617, para.1)	No (p.617, para.1)	Yes (p.617, para.6)
Klinkenberg et al. [2011]	variable (p.1817, para.1)	15 (p.1815, para.1)	No (p.1815, para.1)	No (p.1814, para.6)
Von Davier et al. [2009]	1 year (p.8, para.2)	77-99 (p.8, para.2)	Yes	No
Caughey and Warshaw [2015]	1 year (p.204, para.3)	sparse e.g. 1-2 (p.210, para.2)	Yes	No
Carroll et al. [2016]	1 year (p.70, para.7)	∼80 (p.70, para.7)	Yes	No
Martinez-Plumed et al. [2019]	none	131-625 (p.22, table 1)	Yes	No
Vandemeulebroecke et al. [2017]	6 months (p.636, para.3)	14 (p.636, para.3)	Yes	No
Attali [2011]	none	27 (p.475, para.2)	No (p.475, para.3)	No

Table 1: Parameterization of data collection in the literature

$\mathcal{T}$	Set of all interaction index
t	Interaction time step, $t \in \mathcal{T}$
θ_t	Subject ability at time step t
$\mathcal{Q}$	Item bank
q_t	Item administered at time step t
z_t	Parameters of q_t
y_t	Subject response to q_t (1 for correct and 0 otherwise)
$\mathcal{L}_\tau$	Learning dynamic of subject undergoing feedback
$\mathcal{S}_\psi$	Random change in ability when subject is unobserved
L	Log likelihood of subject's answer

Table 2: Notation

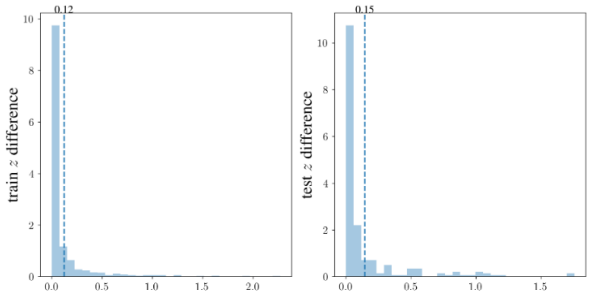

Figure 6: Spread of absolute differences between difficulty of generated item and target difficulty.

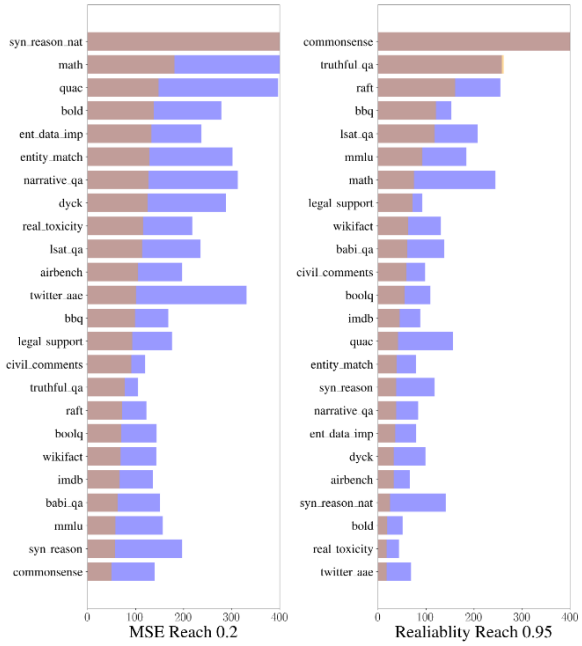

Figure 7: Number of items required to achieve specified accuracy levels across various benchmarks from HELM [Liang et al., 2023] using random sampling (blue) versus adaptive sampling (orange). **Left:** number of items needed until the mean squared error (MSE) of the ability estimate drops below 0.2. **Right:** number of items needed until the marginal reliability of the ability estimate exceeds 0.95, where marginal reliability is defined as $1 - \text{MSE}/\text{Var}(\theta)$. In both panels, adaptive sampling consistently requires fewer items, illustrating the efficiency gains from IRT-based item selection.

A Reference Tables

B Supplementary Material for the Taxonomy

B.1 Supplementary Figures for Static Measurement (Setting 1)

B.2 Gaussian Process Inference for Ability Trajectories

Given a calibrated item bank, starting with an initial parameter ψ of the dynamic model, we can generate a sequence of θ values over time, enabling computing the goodness of fit of the IRT model, which is used for updating ψ. We repeat this process until convergence. We demonstrate the inference process using a simulation. Between time $t_1, ..., t_T$, the ability trajectory $\Theta = [\theta_1, ..., \theta_T]$ is generated as $\Theta \sim \mathcal{GP}(0, \Sigma)$, where $\mathcal{GP}$ is a Gaussian process whose covariance matrix $\Sigma \in \mathbb{R}^{T \times T}$ is constructed using the distance between time steps. At each time step, an item z_t is administered, and the score is generated according to Rasch's model $y_t \sim p(y_t = 1|\theta_t) = \text{Bern}(y|\sigma(\theta_t - z_t))$. Given the observed data $Y = [y_1, ..., y_T]$, the Gaussian process posterior is $p(f|Y) \propto p(y|f)p(f)$. Even though the posterior is not available in close-form, we can still perform inference using MCMC techniques. Leveraging the geometry of the Gaussian process prior to this case, inference can be carried out efficiently with elliptical slice sampling.

Building upon this inference framework, we conducted a simulation study to investigate the impact of spreading measurements over time on the accuracy of the inferred learning trajectory. We varied the number of sampling clusters while keeping the total number of administered items constant. A small number of clusters (with many items per cluster) offers a coarser approximation of the ability's progression (Fig. 8a). Conversely, increasing the number of clusters (with fewer items per cluster) provides a finer temporal resolution of the ability trajectory (Fig. 8b). Improving trajectory inference could lead to better generalization beyond the measured data. This capability would be especially valuable in applications where predictions or decisions need to extend beyond the observed measurement points.

An alternative, computationally simpler solution is an adaptation of the Elo rating system (ERS), originally used to rank chess players. The standard formulation uses a likelihood analogous to the one-parameter logistic function: $P(y = 1|\theta_t, z_t) = (1 + e^{-(\theta_t - z_t)})^{-1}$. Note that ERS used in chess is scaled

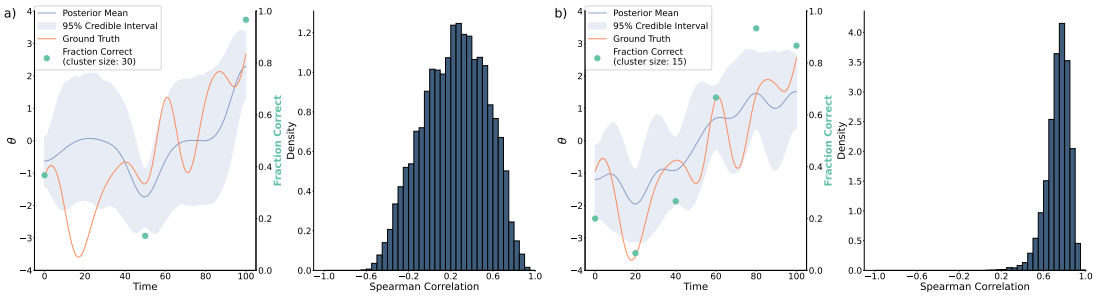

Figure 8: Inference of simulated learning curves under different sampling strategies. a) (Left) Sampling 3 times where each cluster consists of 30 items and (Right) the correlation between the inferred learning rate and the ground truth. b) Sampling 6 times where each cluster contains 15 items.

with a base ten rather than e and divides the exponent by 400 in order to produce scores that are similar to a previous ranking system. Ability and difficulty are updated as $\theta_{t+1} = \theta_t + K(y_t - P(y = 1|\theta_t, z_t)), \phi_{q_{t+1}} = z_t + K(P(y = 1|\theta_t, z_t) - y_t)$, where K is the update weight. The value of K is governed by an uncertainty function, the form of which can vary and may be a function of the number of items answered as well as the timing between measurements [Pelánek, 2016].

C Extended Background on IRT and Knowledge Tracing

C.1 Historical Development of IRT

One of the first models for measurement of a learner's ability comes from classical test theory (CTT), which describes test scores as a function of a true score plus some measurement error [Hambleton and Jones, 1993]. CTT is limited in that ability estimation depends on the test set: the ability estimates of two test-takers are only comparable if they took tests with similar difficulty. The limitations of CTT motivated a series of developments over the 20th century, including factor analysis [Spearman, 1904], generalizability theory [Cronbach et al., 1963], and ultimately item response theory (IRT), each progressively separating the properties of items from the properties of examinees. For a comprehensive review of the different components in measurement experiments, we refer the reader to Scalise and Gifford [2006].

C.2 Computerized Adaptive Testing

In cases where the item parameters z_t are predetermined (from a previous calibration), evaluators can more efficiently ascertain the ability of a subject by adapting item difficulty to the subject's performance during an evaluation session [Van der Linden et al., 2000, van der Linden, 1998]. This is commonly known as computerized adaptive testing (CAT) [Wainer et al., 2000]. Items are typically chosen to maximize Fisher Information: $I(\theta) = E_\theta \frac{\partial^2 L}{\partial \theta^2}$, where E_θ is the expectation with respect to θ and L is the log-likelihood $L = y_t \log P(y_t = 1|\theta, z_t) + (1 - y_t) \log(1 - P(y_t = 1|\theta, z_t))$. When only a few items are administered, the estimate of θ can show a significant bias in the responses to the first few items. Techniques like discrimination stratification [Chang and Ying, 1999] can reduce this bias.

C.3 IRT Applications Beyond Psychometrics

IRT is increasingly useful for measurements of latent variables in many areas beyond psychometrics and learning science. For example, IRT has been applied to neuropsychiatric monitoring [Vandemeulebroecke et al., 2017], where it has been used to assess the progression of cognitive decline in elderly subjects. In political science, IRT models have been employed for measuring political ideologies [Martin and Quinn, 2002], allowing researchers to quantify the ideological positions of Supreme Court justices over time. The marketing field has also benefited from using IRT to analyze consumer responses to advertising and product features [De Jong et al., 2008].

C.4 Knowledge Tracing: Technical Details

In some teaching settings, KT is used to determine the skills (aka knowledge components) that a learner has acquired [Abdelrahman et al., 2023]. Typically, a learner's knowledge is modeled as a latent binary state (either having mastered a skill or not) via a hidden Markov model. In the unlearned state, emissions of a correct answer will be observed with probability $p(\text{guess})$ and will produce an incorrect answer with probability $1 - p(\text{guess})$. In the learned state, an emission of an incorrect answer will be observed with probability $p(\text{slip})$ and a correct answer with probability $1 - p(\text{slip})$. Further, the hidden state can only transition from the unlearned learning state with probability $p(\text{learn})$ (i.e., there is no forgetting). In practice, these parameters are determined by maximum likelihood estimation calculation based on the learner's answers.

KT is closely linked to IRT in that both are attempts to model a learner's (hidden) ability based on responses to items. The two techniques can be used together, where the probability that a learner knows a skill is the IRT 2-parameter logistic model, and all parameters are fitted using MCMC [Xu and Mostow, 2013]. That is, finding the parameters that maximize the posterior probability:

$$\begin{aligned} p(\theta, \phi_2, \phi_3, l, g, s|y) &\propto L(y|g, s, k)\, p(k^{(0)}|\theta, \phi_2, \phi_3) \\ &\times \prod_{t=1}^{T} p(k^{(t)}|k^{(t-1)}, l) \times p(\theta)\, p(\phi_2)\, p(\phi_3)\, p(l)\, p(g)\, p(s), \end{aligned}$$

where l is the transition to the learned state, g is guessing the correct answer in the unlearned state, s is giving the wrong answer in the learned state, and k is the knowledge state. Usually, KT is employed where multiple skills are being evaluated (e.g., if a learner has mastered both addition and subtraction when teaching arithmetic), while IRT is classically used in a univariate setting. However, IRT can also be extended into a multivariate setting [Reckase, 2009, Bonifay, 2019].