Adaptive Assessment at Scale

The Psychometric Science Behind AI-Driven Test Generation

From Item Response Theory to Large Language Models — and what a working implementation taught me

---

The Problem with Measuring Human Knowledge

For most of its history, the Scholastic Assessment Test — taken by nearly two million students each year — allocated the same fixed item set to every candidate regardless of ability level. A student in the 95th percentile spent half the exam on questions that revealed almost nothing about their true proficiency. A struggling student faced items so far above their current ability that each wrong answer increased anxiety without generating useful data.

This is not a failure of test design. It is a fundamental structural limitation of fixed-form assessment — and it is precisely why the SAT itself transitioned to a multistage-adaptive digital format in 2024, joining the GRE, the GMAT and the ASVAB, all of which already run on adaptive delivery.

The psychometric community identified this problem decades ago. The solution — Computer Adaptive Testing (CAT) — has been theoretically mature since the 1970s. The question is no longer whether AI-assisted adaptive assessment works. It is whether the infrastructure to deploy it at scale — cheaply, outside the budgets of the largest testing organizations — now exists.

It does. And the implications for education are substantial.

Part I — The Psychometric Foundation: What Tests Actually Measure

Classical Test Theory and Its Limits

Classical Test Theory (CTT), the framework underlying most 20th-century standardized tests, decomposes an observed score into a true score plus random measurement error. This is intuitive — but mathematically fragile. Item statistics (difficulty, discrimination) are sample-dependent: the “hardness” of an item is defined only relative to the specific population that took the test. A question calibrated on university students looks trivially easy when administered to doctoral candidates.

Item Response Theory: Sample-Invariant Measurement

Item Response Theory, formalized by Georg Rasch (1960) and extended by Frederic Lord (1980), inverts this relationship. Rather than describing items relative to a sample, IRT models the probability of a correct response as a function of the latent ability of the individual and the specific psychometric properties of the item.

The 2-Parameter Logistic Model (2PL):

$$P(\text{correct} \mid \theta, a, b) = 1 / (1 + \exp(-a(\theta - b)))$$

- a — discrimination parameter: how sharply the item distinguishes between nearby ability levels
- b — difficulty parameter: the ability level at which P(correct) = 0.50
- θ — the candidate's latent trait on a standardized scale

The central psychometric property of IRT is parameter invariance (within model constraints): the difficulty of a well-calibrated item does not change when a different cohort takes the test. This property enables comparisons across forms, administrations and populations that CTT cannot support.

Practically, IRT lets you:

1. Build an item bank calibrated to a common latent scale;
2. Compute the Standard Error of Measurement (SEM) as a function of θ — knowing exactly where measurement is precise and where it is not;
3. Select items that maximize Fisher Information at the current ability estimate — the mathematical basis for adaptive testing.

The Information Function: Why Items Are Not Interchangeable

The Fisher Information I(θ) for a 2PL item peaks at θ = b (where the item is most discriminating) and falls off rapidly above and below that point. Administering an item with b = +2.0 to a candidate at θ = −1.0 provides almost no measurement information: they will almost certainly get it wrong, and the wrong answer tells you almost nothing you didn't already know about their ability.

This is the mathematical argument for adaptivity: a test that selects items at the candidate's current θ provides maximum information per item administered.

Part II — Computer Adaptive Testing: 50 Years of Evidence

The CAT Algorithm

The core CAT procedure, refined across decades of operational deployment, is:

1. Administer an initial item near the assumed population mean ($\theta_0 = 0$);
2. Score the response and update θ via Maximum Likelihood Estimation (MLE) or Bayesian Expected A Posteriori (EAP);
3. Select the next item from the calibrated bank that maximizes Fisher Information at the current θ;
4. Apply exposure controls to prevent high-information items from being over-administered;
5. Terminate when SEM(θ) falls below a predefined threshold, or after a maximum number of items.

Documented Efficiency Gains

The efficiency advantage of adaptive over fixed-form tests is one of the most robust findings in psychometrics.

Weiss & Kingsbury (1984) demonstrated that CAT achieves equivalent measurement precision to fixed-form tests using 40–60% fewer items. The intuition is straightforward: a fixed-form test wastes items on both easy questions that high-ability candidates answer trivially and hard questions that low-ability candidates cannot meaningfully engage with. An adaptive test eliminates both tails of wasted information.

Large-scale deployments confirm this at operational scale:

- The GRE moved to a computer-adaptive format in 1992, substantially reducing test length with no degradation in predictive validity for graduate-school performance;
- The GMAT went fully adaptive in 1997; validity studies showed maintained measurement precision at reduced test length;
- The ASVAB has administered adaptive versions (CAT-ASVAB) to hundreds of thousands of military recruits annually, with documented improvements in classification accuracy;
- The SAT Suite completed its transition to a multistage-adaptive digital format in 2024, using IRT to route candidates between modules.

The evidence base is not speculative. Adaptive assessment works at scale.

The Operational Bottleneck: Item Bank Maintenance

The largest practical challenge in running a CAT system is maintaining a calibrated item bank over time. A CAT that repeatedly administers the same high-information items quickly compromises their psychometric integrity through:

- Content exposure: high-ability candidates share item content, producing correlated errors across administrations;
- Item retirement: compromised items must be retired faster than new items complete their calibration cycles;
- Bank depletion: operational CAT programs require item banks 5–10× larger than any single test form to maintain exposure control.

This infrastructure is expensive. ETS, the organization behind the GRE and TOEFL, employs hundreds of item writers and runs continuous calibration field studies; the College Board (which owns the SAT) and ACT operate comparable pipelines for their own programs. Small educational institutions — and the billion-plus students globally who lack access to well-resourced assessment systems — cannot replicate this model.

This is the gap that AI-assisted item generation is positioned to fill.

Part III — Automatic Item Generation: From Templates to Generative Models

The AIG Framework

Systematic use of cognitive models to generate test items — Automatic Item Generation (AIG) — predates large language models substantially. Isaac Bejar (1993) proposed the concept of radial categories for item generation: identifying a prototypical item and generating variants by manipulating surface and deep features within defined cognitive constraints.

Gierl & Haladyna's 2013 volume Automatic Item Generation: Theory and Practice established the canonical framework. The AIG workflow:

1. Cognitive model specification: identify the knowledge components and cognitive operations required to solve a class of items;
2. Item model development: create a template parameterizing surface features (numbers, names, context labels) while holding cognitive demands constant;
3. Algorithmic generation: instantiate the template with sampled parameter values;
4. Psychometric validation: administer a sample of generated items to calibrate the item model.

The key empirical finding from this literature: items generated from the same item model share similar IRT parameter distributions. If you can calibrate the item model rather than individual items, you can predict the approximate difficulty of a newly generated item without individual field testing. This principle dramatically reduces the cost of bank maintenance.

This is not theoretical. The Uniform CPA Exam has incorporated AIG components, and major language-assessment programs use AIG for reading-comprehension item generation at scale.

The LLM Inflection Point

The 2022–2024 period produced rapid empirical evidence on LLM-based item generation. Studies that evaluate ChatGPT-generated multiple-choice items with formal psychometric models — calibrating them with IRT and benchmarking against human-authored items — have found that generated verbal items can be acceptable and comparable to human-written ones on key psychometric properties, while documenting specific shortcomings that still require expert review.

The gap is real but narrowing. In one controlled comparison of reading items administered to several hundred students, LLM-generated and instructor-written items showed similar reliability, with the human items retaining a modest edge. Subsequent work on GPT-4 generation showed that items produced from detailed cognitive-model specifications outperform open-ended generation prompts on discrimination — confirming that structured templates remain important even with frontier models.

The Critical Domain Caveat: Mathematics

Natural-language item generation and mathematical item generation are fundamentally different problems.

In reading comprehension or social-science assessment, an item can be "approximately correct" — a slightly imprecise answer key is recoverable through expert review before

deployment. In mathematics, a wrong answer key is a categorical failure: it invalidates the item completely, and in a CAT system where items may be administered thousands of times before detection, the consequences propagate across every score the item touched.

LLMs are probabilistic text generators. Even frontier models make algebraic errors on a non-trivial fraction of novel mathematics problems. The failure profile is specific:

1. Computational errors: the answer key is incorrect but the problem structure is valid;
2. Constraint violations: the problem is internally inconsistent (parameters produce impossible cases — a “triangle” with sides 1, 1, 10);
3. Distractor collapse: a supposedly incorrect option is actually correct, invalidating multiple-choice scoring.

Pure generative output is insufficient for mathematical items. The required architecture is hybrid:

- Template-based generation providing structural guarantees;
- Symbolic verification via a Computer Algebra System confirming answer correctness;
- Parametric distractor construction producing plausible but provably incorrect alternatives;
- Statistical quality control monitoring item difficulty distribution across generated batches.

Part IV — From Research to Reality: A Working Implementation

This is where theory meets engineering constraints.

The `satlike_problem_creator` (github.com/YouJustMadeTheList/satlike_problem_creator) is an open-source implementation of these principles, built as the generation engine for the Premio Di Nicola — an annual mathematical competition in the style of the SAT, organized by Vincenzo Di Nicola and now in its seventeenth edition. The project demonstrates several design decisions that the academic literature describes but rarely implements end-to-end.

Layer 1 — Symbolic Generation via SymPy

Every problem template in the engine computes its own solution symbolically. A template does not have a hard-coded answer — it evaluates the relevant expression (for instance `sp.Abs(q_expr) / sp.sqrt(m**2 + 1)`) at runtime via SymPy, then formats the result as a LaTeX expression. The answer is correct by construction, not by authorial claim.

This addresses the most dangerous failure mode in AI-assisted item generation: the answer-key error. If the symbolic computation succeeds, the answer is algebraically verified. If it fails (degenerate parameters), the item is never emitted.

Layer 2 — Role-Swapping for Template Entropy

Standard AIG templates produce items with identical structural appearance. Candidates who see several items from the same template quickly recognize the pattern and transfer

procedural knowledge rather than demonstrating conceptual understanding.

The engine addresses this through role-swapping: the same mathematical relationship (the tangency condition d(O, line) = r) generates two structurally distinct question types by choosing at runtime which quantity is unknown — solving for the radius given the intercept, or for the intercept given the radius. The cognitive demand shifts: one version requires isolating r, the other isolating |q|. Both verify correctly via the same underlying formula. This doubles the effective item count per template while preserving the IRT property that items from the same model share parameter distributions.

Layer 3 — Difficulty-Calibrated Parameter Sampling

The `_random_numeric_like` method implements a difficulty-stratified sampling distribution across parameter types:

Type	EASY probability	HARD probability
Integer	80%	25%
Rational	20%	30%
Irrational (√2, π, e)	0%	30%
Symbolic parameter (k, a, n)	0%	15%

This mirrors how human item writers intuitively grade difficulty — hard items use more abstract parameter types that resist numerical intuition. The critical difference is that the distribution is explicit, reproducible, and empirically adjustable rather than left to authorial judgment. The `difficulty_mix` parameter at the exam level controls a weighted probability over {EASY, MEDIUM, HARD} item draws — enabling difficulty profiles ranging from calibration sessions (uniform distribution) to high-stakes competition (front-loaded hard items).

Layer 4 — RAG on Competition Archives

The knowledge-base component ingests historical Premio Di Nicola problem sets (sixteen archived editions, parsing markdown-formatted competition documents) and applies a lightweight structural tag extractor to identify cognitive-pattern frequency: symmetry, functional equations, geometric probability, set logic, number theory. Pattern-frequency analysis across the corpus reveals which structural ideas recur most in competition mathematics — and these frequencies can bias template selection toward historically validated patterns. This is a rudimentary but theoretically sound form of content-distribution matching between generated items and the target competition format.

Layer 5 — Genuine Indeterminacy in Quantitative Comparison Items

The most psychometrically subtle design decision concerns Quantitative Comparison items with correct Answer D (the relationship cannot be determined from the information given). Naively constructed “Answer D” items are common in low-quality test material: they look indeterminate but are actually authoring failures, where the writer forgot to constrain the domain adequately. Experienced candidates learn to spot these and apply elimination heuristics that undermine the item’s discriminating power.

The engine generates genuine Answer D through explicit case-splitting analysis. For the modular-square problem, it verifies via residue-class enumeration that for a given modulus m, the comparison between x^2 mod m and x mod m yields different orderings for different values of x — and only then assigns Answer D. The indeterminacy is constructed, not assumed. This matters for validity: an item whose answer the examinee cannot verify through case analysis is a different cognitive task than one whose indeterminacy is provable, and the IRT parameter profile will differ accordingly.

Output Pipeline and Longitudinal Architecture

Generated exams export to structured LaTeX documents: a test page with questions and a solutions page with step-by-step algebraic explanations. Every generated session is timestamped and archived with full parameter metadata. This archive is the foundation for the next development stage: longitudinal difficulty estimation across administrations. With sufficient data, IRT parameters for each template can be estimated empirically — producing a self-calibrating system whose difficulty model improves as more data accumulates.

Part V — What the Evidence Shows, and What Remains Open

The Established Findings

- Efficiency: CAT achieves equivalent precision in 40–60% fewer items across multiple domains and populations (Weiss & Kingsbury, 1984; Wainer et al., 2000);
- AIG validity: items generated from cognitive models have predictable IRT parameters (Gierl et al., 2012; Embretson, 1999);
- LLM feasibility: for verbal domains, LLM-generated items approach human quality by expert-reviewer ratings and psychometric analysis (multiple studies, 2023–2024);
- Symbolic verification: CAS-verified mathematical item generation eliminates answer-key errors by construction.

Open Research Questions

Long-term exposure effects on AI-generated banks. No longitudinal study has yet established how an AI-generated item bank degrades over time compared to traditionally authored banks. Exposure-control methods were developed for human-written items with sparse bank structures; AI-generated banks may require different algorithms.

Cross-cultural validity. Item contexts (names, scenarios, units) carry cultural assumptions that affect difficulty independent of the underlying construct. Systematic bias analysis of AI-generated mathematical item contexts across populations has not been published at scale.

High-stakes deployment thresholds. The community maintains appropriate caution about replacing field-tested banks with generated ones for high-stakes certification, where the consequences of measurement error are severe and asymmetric. The minimum field-testing sample required to validate a generated item for high-stakes use remains open.

Proof-based mathematics. Current AIG systems, including this one, are restricted to computational and procedural item types — items with unique correct answers verifiable by CAS. Proof-based items (demonstrate that, show why) require a fundamentally different validation architecture that remains an active research problem.

Part VI — The Access Argument

The most compelling case for AI-assisted adaptive assessment is not efficiency. It is access.

ETS, Pearson and ACT spend tens of millions of dollars annually on item development, calibration and bank maintenance. The infrastructure that makes reliable standardized assessment possible is industrialized, expensive, and concentrated in a handful of organizations in wealthy nations.

A student in a well-resourced institution receives years of exposure to calibrated practice materials that give immediate feedback on specific knowledge gaps and adjust difficulty in real time. A student without those resources takes a fixed-form test once a year and receives a single score — with no information about which specific competencies they have mastered and which are underdeveloped. This gap is not a technical problem. It is an economic one.

AI-assisted item generation, combined with IRT-based adaptive delivery, is one of the few developments capable of materially reducing this gap — because it collapses the marginal cost of producing and maintaining a calibrated item bank by one to two orders of magnitude. A pipeline that once implied roughly half a million dollars in item development can be approximated for a few thousand dollars in compute and engineering time, with far lower ongoing maintenance. (These figures are illustrative order-of-magnitude estimates, not audited costs.)

The psychometric theory is mature. The CAT algorithms are decades old and the delivery infrastructure now exists. The symbolic-verification layer is buildable in Python. The remaining work is engineering.

Conclusion

The intersection of Item Response Theory, Computer Adaptive Testing and AI-assisted item generation is not a speculative research agenda. Each component has decades of independent validation. What is new is their convergence into a single deployable stack — and the economic implications of that convergence.

Mathematical item generation remains the hardest technical problem in this stack, precisely because the correctness bar is unforgiving. The answer is not to avoid the problem but to design systems where correctness is verified structurally rather than asserted statistically.

Personalized assessment — where each student’s test is dynamically calibrated to what is known about their specific ability, and where the item bank is continuously refreshed by a symbolically verified generation pipeline — is no longer a research prototype. The architecture exists. The evidence supports it. The question now is deployment.

The implementation referenced here — `satlike_problem_creator` — is available at github.com/YouJustMadeTheList/satlike_problem_creator. Built as the generation engine for the Premio Di Nicola mathematical competition (XVII editions). Contributions welcome.

Davide De Sanctis — Software Developer & Process Engineer
Working at the intersection of AI systems, formal methods, and educational technology.

References

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test Theory for a New Generation of Tests. Lawrence Erlbaum.

College Board (2024). The Digital SAT Suite uses a multistage adaptive design. collegeboard.org. [SAT transition to multistage-adaptive digital format, U.S. spring 2024.]

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. Psychometrika, 64(4), 407–433.

Gierl, M. J., & Haladyna, T. M. (Eds.) (2013). Automatic Item Generation: Theory and Practice. Routledge.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. Medical Education, 46(8), 757–765.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research.

Wainer, H., et al. (2000). Computerized Adaptive Testing: A Primer (2nd ed.). Lawrence Erlbaum.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21(4), 361–375.

On LLM-generated multiple-choice items evaluated with IRT (representative 2023–2024 work): studies benchmarking ChatGPT-generated reading-comprehension and domain items against human-authored items using psychometric models, reporting comparable-but-imperfect item quality requiring expert review.