From Item Generation to Agentic Measurement Systems: A Conceptual Framework for the Validity of AI-Driven Assessments

Alina A. von Davier[1,*,†]

[1]Duolingo

Abstract

The rapid adoption of large language models (LLMs) in assessment has shifted the focus from isolated item generation to designing complex, artificial intelligence (AI)-driven measurement systems. This paper describes an agentic AI framework with human governance for scalable, valid, and transparent assessment development. The approach leverages three collaborating agents, a Task Agent that interprets assessment blueprints, a Generator Agent that produces test content, and a Reviewer Agent that evaluates quality, embedded within a continuous human-in-the-loop and audit-driven ecosystem. Moving beyond traditional automation, this framework reconceptualizes assessment as a dynamic process in which validity, fairness, and reliability are distributed across interacting agents and human experts. Drawing on examples from operational testing contexts, the paper highlights the need to reconsider evaluation and validity for agent-based assessment systems. It also explores how process data, audit trails, and iterative feedback loops can serve as new sources of psychometric evidence. The paper concludes by emphasizing the need to evaluate not only the automatically generated items but also the agents' behavior and their reliability.

Keywords

Assessment, Validity, Agentic AI, Test development, Computational psychometrics

1. Introduction

Advances in large language models (LLMs) have accelerated the adoption of artificial intelligence (AI)in educational assessment, particularly in the automation of item generation. While early applications focused on improving efficiency in content production, current developments signal a broader transformation: assessment systems are evolving from collections of static, human-authored items into dynamic, AI-driven measurement ecosystems. This transformation is driven by the increasing need for scale, diversity, and responsiveness in digital, adaptive assessments. Traditional item development processes, which rely heavily on expert item writers, are resource-intensive and difficult to scale. At the same time, the generation of high-quality items remains cognitively demanding and requires careful alignment with constructs, linguistic precision, and fairness considerations [1]. Fully manual approaches cannot meet modern demands, while naïve automation risks compromising validity and control.

This paper argues that addressing this tension requires moving beyond task-level automation toward system-level design. I describe agentic measurement systems, in which multiple AI agents collaborate under human governance to produce, evaluate, and refine assessment content. I position this work as a theoretical and design contribution to the learning analytics community, focusing on how such systems can be conceptualized, designed, and evaluated.

LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners, LAK 2026 Workshop

*Paper based on the keynote presented at the LLM Psychometrics Workshop at LAK 2026.

Duolingo, 5900 Penn Avenue, Pittsburgh, PA 15206, USA

avondavier@duolingo.com (A. A. von Davier)

0000-0002-5505-5311 (A. A. von Davier)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-15

CEUR-WS.org/Vol-4330/paper9.pdf

2. Background and Motivation

Assessment has traditionally been guided by individual theoretical frameworks, including the validity argument framework [2, 3], the assessment use argument framework [4], and established professional standards in educational measurement (e.g., [5]). While foundational, these frameworks either do not explicitly address the role of AI in assessment or consider it only as a tool for automation in a limited and fragmented way (for example, the automated scoring models have been included in the validity frameworks). Moreover, they tend to treat key components of assessment—such as design, delivery, scoring, and evaluation—as largely independent processes, without fully accounting for the complex interactions among them.

In contrast, the ecosystem model proposed by [6] explicitly captures the interdependencies across the full range of assessment processes. For example, it highlights how design decisions, particularly those involving AI—shape measurement approaches, and how principles of responsible AI influence practices across domains such as test development, scoring, and security.

The present paper adopts this dynamic ecosystem approach to assessment. Compared to traditional, standalone frameworks, an ecosystem perspective provides a more comprehensive lens for understanding the interconnected nature of modern, AI-enabled, digital-first assessments—tests that are conceived, developed, delivered, and maintained through integrated, AI-driven processes.

AI-supported assessment ecosystems are commonly implemented as structured pipelines that integrate generation, review, and quality assurance processes. These "item factory" approaches combine automation, machine learning, and human expertise to accelerate item production while maintaining quality [7]. However, they still are largely linear and rely heavily on human intervention to ensure validity, and they need to be updated to account for a dynamic production.

Recent advances in AI architectures suggest a different paradigm. Rather than treating AI as a tool executing static prompts, agent-based approaches conceptualize AI as a set of interacting components that collaborate within a shared environment. This shift moves from automation to agency and from tools to systems as shown in [8].

From a learning analytics perspective, this transition aligns with broader efforts to model complex educational systems in which outcomes emerge from interactions among components. It also raises new questions about how to represent constructs, evaluate processes, and ensure accountability in AI-driven systems.

3. A Conceptual Model of Agentic Measurement Systems

In this section I describe the agentic measurement systems as multi-agent architectures embedded within a human governance layer as introduced in the upcoming book chapter [8] where implementation details are also provided. The model consists of three primary agents that collectively operationalize assessment development. See Figure 1 for a description of the process.

The Task Agent translates theoretical constructs and assessment frameworks into structured specifications. The Task Agent functions as an intermediary between the psychometric intentions of assessment developers, who rely on frameworks such as Evidence-Centered Design (ECD, [1]) and its extension, eECD ([9]), and the AI-driven content generation process. Its central responsibility is to convert test specifications, including construct definitions, targeted difficulty levels, and content constraints, into structured, machine-interpretable instructions that downstream agents can act upon. In doing so, the Task Agent effectively replaces the manual interpretation of test blueprints with an automated process that remains transparent and traceable. By encoding psychometric intent in machine-interpretable formats, it bridges the gap between assessment theory and computational processes.

The Generator Agent assumes responsibility for producing the assessment content, once the specifications have been defined. This component, powered by large language models (LLMs), generates the actual test materials–such as reading passages, prompts, questions, and response options–based on the structured input provided by the Task Agent. In this sense, it performs a role similar to that of a human

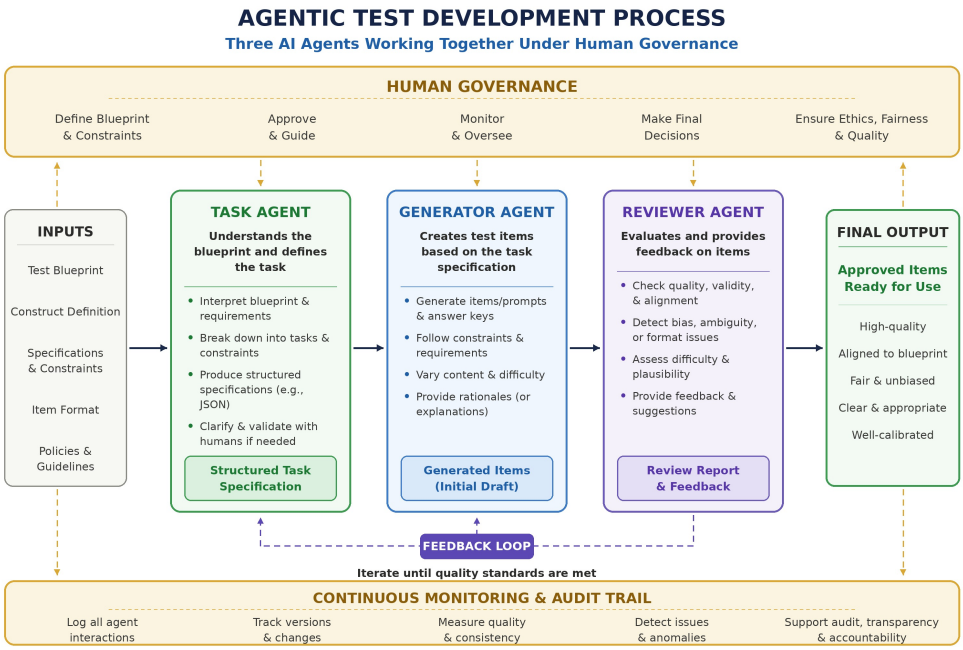


Figure 1: Visual representation of the agentic AI for test development. Figure created with Claude Fable 5.

item writer, but with the added advantage of scalability and the ability to produce a diverse range of outputs.

The Generator Agent may consist of one or multiple LLMs, each assigned a specialized function within the broader agentic system. It takes structured inputs (for example, JSON format), and follows role-specific system prompts that ensure adherence to all relevant item specifications and framework constraints while producing outputs in a consistent, standardized format. To further guide generation, few-shot examples of high-quality items can be included, helping to shape aspects such as tone, structure, and linguistic complexity.

The Reviewer Agent serves as the system's internal evaluator, assessing generated content against criteria related to fairness, content (e.g., linguistic complexity, topical appropriateness), and psychometric alignment. See [10] for details on the fairness and bias review for the Duolingo English Test.

Like other components, it can be implemented using one or more LLMs with specialized role definitions, supported by evaluation rubrics, bias-detection tools, and the ability to invoke external scripts or functions. In the future, these components may become sub-agents, depending on the degree of autonomous behavior and complexity of decision making situation.

A key strength of LLM-based agents is their capacity to determine when to rely on external tools. For instance, rather than performing tasks they are not well-suited for, such as accurately counting words, the Reviewer Agent can call a dedicated function and incorporate its output into the evaluation process. This integration enables more reliable and efficient assessment workflows. At the moment this capacity does not always function well, or it may become problematic, if the agents have too much access to the computer or drive. This will be discussed in more details the next section.

The Reviewer Agent identifies issues such as construct-irrelevant content, cultural bias, or redundancy, and can request revisions from the Generator Agent. This establishes an iterative refinement loop in which both agents collaborate autonomously until predefined quality thresholds are satisfied or the maximum number of iteration was achieved. For example, the Reviewer Agent may be initialized and configured with quality criteria, evaluation metrics, and the original task specifications, which it uses to assess elements such as distractors in an item. It then returns structured feedback to the Generator Agent, along with the original content, prompting revisions. This cycle continues until the content meets all criteria or a maximum number of iterations is reached, at which point the item may be discarded.

These agents interact through iterative refinement loops, in which outputs are continuously evaluated and improved. The Reviewer Agent evaluates the output, provides feedback to the Generator Agent, who regenerates the output.

3.1. Human Governance

Human experts define constraints, monitor system behavior, and intervene when necessary, forming a governance layer that ensures alignment with theoretical and ethical standards. The human-in-the-loop practices function both as a safeguard and as a calibration mechanism, ensuring that agent-generated outputs remain aligned with theoretical frameworks and empirical expectations. In practice, human experts, such as assessment specialists, psychometricians, test developers, and AI engineers can intervene at multiple critical stages of the process. For example, during initial development and fine-tuning, experts evaluate how the Task Agent interprets key blueprint components, including construct definitions.

At later stages, human reviewers may also act as real-time evaluators of AI-generated content. Subject matter experts (SMEs), for instance, can intervene within the interaction loop between the Generator and Reviewer Agents, particularly when new item types are introduced. In such cases, the Generator Agent can produce multiple task variants in parallel, enabling efficient piloting and comparative evaluation by human reviewers.

3.2. Agentic AI for Psychometrics

Although this paper focuses on the use of agentic AI for test development, parallel efforts are underway to extend similar principles to psychometrics. Emerging work in this area suggests promising directions, drawing on advances such as Google's Co-Scientist and related systems. These developments offer valuable insights that can inform the design and implementation of agentic approaches in psychometric modeling and analysis.

One could envision a *DataAnalyst* AI agent, that checks and cleans the data; a *PsychometricAnalysist* AI agent who would run a specific model (or multiple models and choose the best model according to various criteria), and a *PsychometricReviewer* AI agent, that evaluates the results. Specific design principles for psychometric work and quality control metrics will need to be developed for these types of agents.

Once the set of AI agents for the psychometric work is developed and appropriate integration and control mechanisms are put in place, then, the (modular) *agenticmeasurementsystem* will be complete. Obviously, significant care will be needed in setting that work as well.

In principle, it is possible to consider deploying these agents in real time and deliver items (adaptively) to test takers. More research is necessary for monitoring and controlling agent behavior.

4. Design Principles

Based on assessment science and engineering best practices, key design principles are identified for agentic assessment systems. These design principles are also aligned with the evidence centered design [1]. First, the construct definition and the task design need to be explicitly stated and the Task Agent needs to have retrieval capabilities and access to the relevant information. Second, guidelines and rubrics for ensuring fairness of the generated content needs to be provided. Third, psychometric intent must be explicitly encoded in structured representations. This ensures that generative processes remain aligned with constructs and specifications. Fourth, responsible AI guidelines need to be provided. Fifth, systems should be designed around iterative feedback loops. Continuous cycles of generation and evaluation enable adaptive improvement and produce process data that can be used for monitoring and validation. Sixth, human oversight should be continuous and integrated throughout the system. Rather than acting solely as final reviewers, human experts define constraints, monitor intermediate outputs, and guide system behavior. Seventh, transparency and auditability are essential. Systems should

produce structured outputs and maintain logs of decisions, enabling inspection and accountability. Audit trails and process data are critical components of quality assurance. Finally, modularity allows systems to evolve and to be controlled by the hu,man experts. Additional agents can be introduced to manage orchestration, calibration, or optimization as system complexity increases.

4.1. Safe Configuration for Agentic Workflows

Since these agents work autonomously and at high speeds, strict protocols are necessary to protect our data and processes.

Here are several key recommendations:

- Strict sandboxing: AI agents should never interact directly with the host machine's file system or terminal. They must be isolated in containerized environments, such as Docker.
- Verification protocols: All AI-generated instructions need to be treated as untrusted. One should not grant agents blanket API permissions, root access, or write access to production databases.
- Human governance and human-in-the-loop practices: One should configure tools so that any destructive action—such as bulk deletions, directory wipes, or modifications to database items, item pools, and test-taker data—requires manual human approval. Additionally, alarms and automatic system shutdowns need to be implemented if such actions are attempted without authorization.
- External backups: External backups need to be maintained and code needs to be pushed frequently to remote version control repositories to ensure that a local catastrophe does not result in permanent data loss.

The autonomy and speed of these agents present significant risks. Without these safeguards, we face potential consequences for test validity, including the risk of item pools being erased or modified and the loss of test-taker data, which would prevent scores from being sent to institutions.

5. Validity as a System Level Characteristic

The emergence of agentic systems challenges traditional conceptions of validity by shifting attention from static outcomes to dynamic, system-level processes. In conventional frameworks, validity is primarily concerned with the interpretation and use of test scores in relation to underlying constructs [2, 3]. In traditional test development settings, the policies that support the quality of work and the behavior of human experts are managed externally by institutions and test providers. However, in agentic systems, where content generation, scoring, and evaluation are mediated by interacting AI components, the processes that produce both items and scores become inseparable from the validity argument itself. While the institutions' policies and reputation continues to matter, the management of humans needs to extend to AI agents.

In this context, validity can be understood as distributed across system components rather than residing solely in score interpretations. Each agent contributes distinct evidence to the overall validity claim: the Task Agent supports construct representation by translating theoretical specifications into operational definitions; the Generator Agent contributes to evidence elicitation by producing tasks aligned with those specifications; and the Reviewer Agent ensures quality control through evaluation of fairness, alignment, and psychometric integrity. Human governance further anchors the system by providing oversight, adjudicating ambiguous cases, and ensuring alignment with theoretical principles, ethical standards, and intended use.

This perspective extends validity beyond outcomes to encompass the behavior and interactions of the system as a whole [6]. It aligns with principles from computerized system validation, where system reliability depends not only on outputs but on the verification, traceability, and control of underlying processes. In such systems, validation requires documented evidence that each component performs

as intended, that interactions between components are well understood, and that the system remains stable under operational conditions.

Accordingly, agentic assessment systems introduce new challenges for validity theory and practice. These include evaluating the reliability and consistency of individual agents, detecting and managing model drift over time, ensuring reproducibility of generated content, and integrating process-level evidence (such as logs, decision traces, and tool usage) into validity arguments. Hence,this perspective extends validity beyond outcomes to include processes and agents' behavior. Addressing these challenges requires expanding traditional validity frameworks to incorporate continuous monitoring, system auditing, and lifecycle validation, recognizing that in AI-enabled environments, validity is not a one-time judgment but an ongoing, evidence-based process.

5.1. Process Data, Learning Analytics, and Computational Psychometrics

Agentic systems generate rich forms of process data, including agent decisions, intermediate outputs, revision histories, tool calls, and comprehensive audit trails. These data extend far beyond traditional assessment outputs by capturing how results are produced, not just what the results are. As such, they open new opportunities for analysis, evaluation, and system improvement.

From a learning analytics perspective, these data can be understood as traces of activity within a complex socio-technical system. Learning analytics is broadly concerned with the measurement, collection, analysis, and reporting of data about learners and learning environments in order to understand and optimize processes and outcomes. In the context of agentic assessment systems, this perspective can be extended to include not only human learners and test takers, but also AI agents as participants in a dynamic system. The interaction logs and decision traces generated by agents can therefore be modeled as sequences of actions, enabling the study of system behavior, adaptation, and performance over time [11]. For instance, systematic patterns in Reviewer Agent feedback may reveal latent biases or inconsistencies in evaluation criteria, while the number of iterations required for convergence in Generator–Reviewer loops may serve as an indicator of system efficiency, stability, or task complexity.

Integrating such process data into validation frameworks represents a significant shift from a sole focus on product-based evidence (e.g., test scores) toward a more comprehensive approach that combines both process and outcome evidence. This perspective is closely aligned with developments in computational psychometrics [12], an emerging field that integrates psychometric theory with computational methods to model complex, data-rich assessment environments. Computational psychometrics emphasizes the use of fine-grained data, such as response processes, interaction logs, and system behaviors, to better understand both measurement constructs and the mechanisms through which evidence is generated. Within agentic systems, this means that validity arguments can be strengthened by incorporating evidence about how items are generated, evaluated, and refined, not just how they function statistically after deployment.

Together, learning analytics and computational psychometrics provide complementary lenses for analyzing agentic assessment systems: the former focuses on understanding and optimizing system dynamics through trace data, while the latter provides the methodological and theoretical tools to incorporate such data into rigorous measurement models. This integration highlights the growing importance of process-oriented evidence in the design, evaluation, and validation of AI-driven assessment systems.

6. Governance, Responsible AI and Ethics

The deployment of agentic systems requires robust governance frameworks and subsystems. Ensuring validity, fairness, protection, and transparency of a high-stakes test and its development involves integrating responsible AI principles throughout the system lifecycle. See [13, 14, 15]for examples of this type of integration for the Duolingo English Test. Key mechanisms include institutional guidelines, regular audits, and compliance with data privacy and security standards. Human oversight remains central, particularly in high-stakes contexts. Responsible AI in this context extends beyond evaluating

outputs to monitoring agent behavior and system interactions. This requires new forms of accountability, including the ability to audit decision processes and explain system outputs. Ethics need to be considered around the maintenance and support of human values [16].

Even when well-conceived and carefully implemented, agentic AI systems for assessments present numerous risks and vulnerabilities, ranging from malfunctions and hallucinations to information erasure. While research and development in this area is still considered nascent, it is improving rapidly as the community continues to develop best practices.

7. Discussion and Future Work

Agentic measurement systems represent a shift toward dynamic, process-oriented assessment. While they offer significant advantages in scalability and adaptability, they also introduce new challenges in evaluation and control. Future research should focus on evaluating agents as components of measurement systems, including their reliability, consistency, and interaction effects. Additional work is needed to understand how system behavior evolves over time and how it can be monitored and controlled. The integration of agentic systems with adaptive testing and learning environments also presents opportunities for transforming assessment into a continuous process.

This paper has described a theoretical and design framework for agentic measurement systems, positioning them as the next stage in AI-driven assessment. By integrating multiple AI agents with human governance, these systems enable scalable and adaptive assessment development while maintaining a focus on validity and fairness [2, 13].

This paper posits that the AI capabilities, especially agentic work, requires rethinking validity as a system-level property and developing new methods for evaluating both outputs and processes of test development. This proposal for a system-level validity resonates with the dynamic ecosystem perspective as described in [6], but it would require additional components for the management of agents' behavior.

This paper posits that realizing this vision requires rethinking validity as a system-level property and developing new methods for evaluating both outputs and processes of test development. This proposal for a system-level validity resonates with the dynamic ecosystem perspective as described in [6], but it would require additional components for the management of agents' behavior.

Ultimately, the future of assessment lies in designing and governing intelligent systems that produce meaningful and trustworthy measures of human ability.

Acknowledgments

The author has no conflict of interest. The author thanks the anonymous reviewers for their comments.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT Plus in order to summarize information, improve grammar and edit, and reformat the references. Claude Fable 5 was used to create Figure 1. After using these tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] R. J. Mislevy, R. G. Almond, L. A. Lukas, A Brief Introduction to Evidence-Centered Design, Technical Report RR-03-16, Educational Testing Service (ETS), Princeton, NJ, 2003.

[2] M. T. Kane, An argument-based approach to validity, Psychological Bulletin 112 (1992) 527–535. doi:10.1037/0033-2909.112.3.527.

[3] C. A. Chapelle, M. K. Enright, J. M. Jamieson (Eds.), Building a validity argument for the test of English as a foreign language, Routledge, 2008. doi:10.4324/9780203937891.

[4] L. F. Bachman, A. Palmer, Language assessment in practice, Oxford University Press, 2010.

[5] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Standards for educational and psychological testing, American Educational Research Association, 2014.

[6] J. Burstein, G. T. LaFlair, A. A. von Davier, A theoretical assessment ecosystem for a digital-first assessment—the Duolingo English Test, Technical Report DRR-25-05, Duolingo, 2025. URL: https://go.duolingo.com/ecosystem.

[7] A. A. von Davier, A. Runge, Y. Park, Y. Attali, J. Church, G. T. LaFlair, The item factory, in: H. Jiao, R. W. Lissitz (Eds.), Machine learning, natural language processing, and psychometrics, Emerald Publishing Limited, Leeds, England, 2024, pp. 10–42.

[8] A. A. von Davier, Agentic AI with human governance for assessment content generation, in: A. A. von Davier, D. Yan (Eds.), Artificial intelligence in educational learning and assessment, Springer, 2026.

[9] M. Arieli-Attali, S. Ward, J. Thomas, B. Deonovic, A. A. von Davier, The expanded evidence-centered design (e-ECD) for learning and assessment systems: a framework for incorporating learning goals and processes within assessment design, Frontiers in Psychology 10 (2019) 853. doi:10.3389/fpsyg.2019.00853.

[10] J. Church, Y. Park, J. Burstein, Guidelines for fair test content: the Duolingo English Test example, Technical Report DRR-25-02, Duolingo, 2025. URL: https://go.duolingo.com/DETFairnessGuidelines.

[11] P. Chopade, K. Stoeffler, S. M. Khan, Y. Rosen, S. Swartz, A. A. von Davier, Human-agent assessment: Interaction and sub-skills scoring for collaborative problem solving, in: Artificial Intelligence in Education: 19th International Conference, AIED 2018, Proceedings, Part II, volume 10948 of *Lecture Notes in Computer Science*, Springer, Cham, 2018, pp. 52–57. doi:10.1007/978-3-319-93846-2_10.

[12] A. A. von Davier, R. J. Mislevy, J. Hao (Eds.), Computational psychometrics: new methodologies for a new generation of digital learning and assessment: with examples in R and Python, Springer International Publishing, 2022. doi:10.1007/978-3-030-74394-9.

[13] J. Burstein, The Duolingo English Test Responsible AI Standards, Technical Report DRR-25-05, Duolingo, 2025. URL: https://go.duolingo.com/ResponsibleAI.

[14] J. Burstein, G. T. LaFlair, K. Yancey, A. A. von Davier, R. Dotan, Responsible AI for test equity and quality: the Duolingo English Test as a case study, arXiv preprint (2024). doi:10.48550/arXiv.2409.07476.

[15] J. Burstein, A. A. von Davier, N. Shaked, Custodianship and responsible AI in educational learning and assessment, in: A. A. von Davier, D. Yan (Eds.), Artificial intelligence in educational learning and assessment, Springer, 2026.

[16] A. A. von Davier, J. Burstein, Ai in the assessment ecosystem: Implications for fairness, bias, and equity, in: P. Ilic, I. Casebourne, R. Wegerof (Eds.), Artificial Intelligence in Education: The Intersection of Technology and Pedagogy, Springer Nature Switzerland, 2024.