Assessment in the Age of AI

Feedback, Scoring, and Validity in Higher Education

Emrah EMİRTEKİN

PEGEM AKADEMİ

Emrah EMİRTEKİN

ASSESSMENT IN THE AGE OF AI
Feedback, Scoring, and Validity in Higher Education

ISBN 978-625-8897-38-8

Pegem Academy Publishing Training and Consultancy Services Tic. Ltd. Şti.It is a publishing house founded in 1998 in Ankara/Turkey which has been operating regularly for 27 years. Published books, it is included in the catalogs of higher education institutions. Pegem Academy has over 2000 publications in the same field from different authors. Detailed information about our publications can be found http://pegem.net.

1st Edition: September 2026, Ankara

Publication-Project: Selcan Özgürel
Typesetter-Graphic Designer: Beyza Nur Erdoğan
Cover Design: Pegem Akademi

Printed by: Sayfa Basım Sanayi Ticaret Ltd. Şti.
İvedik OSB Matbaacılar Site 1514. Street No: 23-25
Yenimahalle/ANKARA

Publishing House Certificate Number: 51818
Printing House Certificate Number: 77079

Contact

Pegem Akademi: Shira Trade Center
Macun Neighbourhood 204th Street Number: 141/A-33 Yenimahalle/ANKARA
Publishig House: 0312 430 67 50
Distribution: 0312 434 54 24
Preparatory Courses: 0312 419 05 60
Internet: www.pegem.net
E-mail: yayinevi@pegem.net
WhatsApp Line: 0538 594 92 40

FOREWORD

There are moments in an academic life when we have the privilege of watching a young scholar gradually find not only a field of research, but also a question that is genuinely his or her own. For me, Emrah Emirtekin's academic journey has been one of those experiences.

I first came to know Emrah as a student at the undergraduate level and later had the opportunity to accompany him through his graduate education as an academic advisor. Over the years, our paths continued to intersect through our shared work in educational technologies and, for nearly a decade, through the activities of the Yaşar University Open and Distance Learning Research Centre. This has allowed me to observe his development not simply as a student who became a researcher, but as a scholar who gradually learned to formulate increasingly important questions about technology, learning, evidence, and educational practice.

This book, Assessment in the Age of AI: Feedback, Scoring, and Validity in Higher Education, is, in many ways, a reflection of that journey. It is also a timely response to one of the most consequential questions currently facing higher education: what should assessment mean when artificial intelligence can generate feedback, assign scores, explain decisions, and increasingly participate in processes that were once regarded as exclusively human?

What I particularly value in Emrah's work is that he does not approach this question with either technological enthusiasm or technological scepticism. Instead, he asks a much more difficult and much more useful question: under what conditions can AI genuinely contribute to assessment without compromising learning, fairness, validity, or human responsibility? This is an important distinction. The book begins from the recognition that the fact that an AI system can produce fluent feedback or scores that resemble those of human raters is not, by itself, sufficient evidence that the system is educationally trustworthy.

That position gives the book its intellectual character.

The work is organized around three closely connected concerns: designing feedback, governing scoring, and validating the evidence behind both. The first part examines AI-supported feedback through the lenses of feedback literacy and human-AI collaboration. Rather than presenting AI as a replacement for teachers, Emrah positions it as one participant in a broader feedback ecosystem in which students, teachers, peers, and AI systems have different responsibilities. His proposed Feedback-Literate Hybrid Intelligent Feedback (FHIF) framework is particularly meaningful in this regard because it places student agency, explainability, accountability, and human judgment alongside the technological capabilities of AI.

The second part moves from pedagogy to institutional responsibility. Here, the discussion of LLM-based scoring goes beyond the familiar question of whether an AI system agrees with a human rater. Privacy, cybersecurity, fairness, explainability, auditability, teacher agency, and human oversight become integral parts of the assessment problem. This is a crucial contribution because an assessment system is never merely a technical system. It is also a social, institutional, ethical, and human system. The book therefore treats LLM-based scoring as socio-technical assessment infrastructure rather than simply as another automated scoring tool.

The third part brings the discussion to a question that I believe is especially important for researchers in educational technology: How do we know that the claims we are making about AI-based assessment are actually justified? The chapter on validity provides a methodological

bridge between promising technological demonstrations and defensible educational research. It emphasizes reliability, bias, reproducibility, study design, and reporting standards, reminding us that an impressive agreement statistic is evidence, not permission for adoption.

This methodological caution is one of the strongest features of the book. Emrah does not claim to have produced a new scoring model or a universally applicable benchmark. Nor does he present the frameworks developed in the book as empirically validated solutions. Instead, the work deliberately offers conceptual, methodological, and governance-oriented tools for researchers and practitioners who need to make better decisions about AI in assessment.

The book is also notable for the way it brings different traditions of educational technology research into conversation with one another. Assessment, learning sciences, feedback literacy, artificial intelligence, learning analytics, measurement, human-AI collaboration, privacy, cybersecurity, and educational governance are often studied in separate scholarly conversations. Emrah's work attempts to bring these perspectives together around a common educational problem. In doing so, he reminds us that the question is not simply whether an AI model is accurate, but whether its use is educationally meaningful, methodologically defensible, ethically responsible, and institutionally sustainable.

I have also followed Emrah's development as a researcher with particular interest because the questions addressed in this book did not emerge in isolation. His recent scholarship includes systematic work on LLM-powered automated assessment and empirical research examining AI-human agreement in short-answer grading. These studies provide an important scholarly foundation for the arguments developed here. His academic profile similarly reflects a growing focus on artificial intelligence, automated assessment, learning analytics, and data-driven educational systems.

For me, however, the significance of this book goes beyond the individual studies and frameworks it brings together. It represents a stage in an academic journey that I have had the opportunity to witness from relatively close quarters. I have seen Emrah move from being a student who was learning how to ask research questions to becoming a researcher who is now asking questions that the field itself needs to confront.

That transition is not always visible in a publication. It is visible here.

There is a maturity in the central argument of this book: technological capability should not be confused with educational value, and technical performance should not be confused with validity. In an era in which AI systems are developing faster than many of our educational policies, assessment practices, and research methodologies, this distinction is more important than ever.

I believe this is also where the book can make a meaningful contribution to the field. For researchers, it offers a conceptual and methodological lens through which emerging AI assessment studies can be designed and interpreted more carefully. For instructors and academic developers, it provides practical ways of thinking about when AI may support feedback and assessment and when human judgment must remain central. For institutions, it draws attention to governance issues that cannot be solved simply by selecting a more capable model. And for scholars entering this rapidly developing field, it offers something perhaps even more valuable: a way of framing the questions that should be asked before we rush to implement the answers.

The book's message is ultimately not anti-AI, nor is it a celebration of automation. It is a call for responsible human-AI collaboration in education. AI can expand opportunities for timely feedback, support revision, reduce repetitive workload, and make assessment criteria more visible. Yet it should not obscure responsibility, replace meaningful human relationships, or become an unquestioned authority in high-stakes educational decisions. The book's first chapter captures this orientation particularly well in its argument for feedback-literate human-AI collaboration rather than fully automated feedback.

I hope readers will therefore approach this book not as a manual for adopting the next AI assessment tool, but as an invitation to think more carefully about assessment itself.

The questions raised here will remain with us long after particular models, platforms, and prompting techniques have changed. How should feedback be designed so that students actually learn from it? What does it mean for an AI-supported score to be fair? What evidence is sufficient to trust an automated judgment? Where should human responsibility begin and end? How should students be able to question or contest an AI-influenced assessment? And, perhaps most importantly, how can we ensure that technological progress remains connected to the educational purposes for which technology is introduced in the first place?

These are not easy questions. They should not be.

It gives me particular pleasure, therefore, to introduce this work by a scholar whose academic development I have had the opportunity to witness over many years. I am proud to see Emrah Emirtekin now contributing to a field in which the questions are becoming increasingly urgent and increasingly consequential.

I believe Assessment in the Age of AI will be valuable not because it claims to have settled the debate, but because it helps us have a better one. It provides researchers with questions worth investigating, practitioners with principles worth considering, and institutions with issues that cannot responsibly be postponed.

Most importantly, I believe this book will serve as a useful guide for those who are trying to understand not merely what AI can do in assessment, but what we should allow it to do, and why.

I warmly recommend this work to researchers, educators, instructional designers, assessment specialists, and educational technology scholars who are navigating the rapidly changing relationship between artificial intelligence and higher education.

It has been a privilege to witness the academic journey that led to this book. I am confident that this work will become an important step not only in Emrah's scholarly development, but also in the continuing conversation about trustworthy, human-centered, and evidence-informed educational technology.

Prof. Dr. Yasin Özarslan

Yaşar University, Türkiye, 2026

PREFACE

This book began with a practical worry rather than a grand claim. As large language models entered higher education, it became easy to show that a model could produce fluent feedback, or a score that resembled a human rater's, and just as easy to treat that resemblance as proof that the technology was ready to use. The chapters collected here grew out of the conviction that resemblance is not enough, and that assessment in the age of AI has to be approached as a human-centered socio-technical practice rather than an accuracy problem to be solved by larger models.

The book is organized around three questions that follow from that conviction. The first chapter asks how AI-generated feedback should be designed so that students actually learn from it, treating feedback as part of a human-AI ecosystem rather than a substitute for teacher judgment. The second asks how institutions can adopt LLM-based scoring responsibly, taking privacy, security, fairness, and human oversight as obligations that accuracy alone cannot satisfy. The third asks how the claims underlying both, that a model's feedback or scores are good enough to build on, can be validated through defensible study design, appropriate reliability and bias evidence, and reproducible reporting. Read together, the three chapters trace a single arc: designing feedback, governing scoring, and validating the evidence beneath both.

The book is written for researchers, instructors, and academic developers in educational technology and assessment who are weighing what these systems can and cannot responsibly do. It does not offer a finished model or a benchmark to adopt. It offers a way of thinking, and a set of design, governance, and validation habits, for using generative AI in assessment without losing sight of the students and teachers it is meant to serve.

Emrah Emirtekin

Manisa, Türkiye

OVERVIEW

Generative artificial intelligence, and large language models (LLMs) in particular, has moved quickly from demonstration to deployment in higher education, where models now draft feedback, assign scores to essays and short answers, and summarize patterns across large cohorts. A recurring pattern accompanies this shift: a system is shown to produce feedback that reads well, or scores that resemble a human rater's, and that resemblance is treated as sufficient reason to adopt it. This book begins from a different premise. Resemblance is a starting point rather than a conclusion, and assessment in the age of AI is best understood not as an accuracy problem to be solved by larger models, but as a human-centered socio-technical practice in which how feedback is designed, how scoring is governed, and how claims are validated matter as much as how closely a model imitates a human.

The three chapters develop this premise along a single arc: design, govern, validate. The first concerns feedback, asking how AI-generated feedback should be designed so that students actually learn from it, with the model treated as one source in a human-AI ecosystem rather than a replacement for teacher or peer judgment. The second concerns scoring, asking how institutions can adopt LLM-based scoring responsibly, with privacy, security, fairness, and human oversight treated as obligations that accuracy alone cannot discharge. The third concerns validity, asking how the claims on which the first two depend, that a model's feedback or scores are good enough to build on, can be established through defensible study design, appropriate reliability and bias evidence, and reproducible reporting. Feedback design and governance both presuppose trustworthy scores and comments; the validity chapter supplies the methods by which that trust is earned.

A brief note on terminology. AI feedback refers to comments or suggestions generated or supported by an AI system to help a learner revise; LLM-based scoring refers to workflows in which a large language model assigns or recommends a score for a constructed response; and validity refers not to any single statistic but to the evidence-supported argument that scores or feedback mean what they are claimed to mean for a given population and purpose. These activities overlap in practice, but raise distinct design, governance, and measurement questions, and the book keeps them separate for that reason.

The book is written for researchers, instructors, and academic developers who must decide what these systems can and cannot responsibly do. Its aims are bounded: it proposes no new scoring model, prompting method, or benchmark, and it does not claim that its frameworks have been empirically validated; the frameworks in the first two chapters are offered as design and governance heuristics, and the third chapter is explicit that such claims require the evidence it describes. What it offers is a way of thinking, and a set of design, governance, and validation habits, for using generative AI in assessment without losing sight of the students and teachers it is meant to serve.

CONTENTS

CHAPTER 3

VALIDITY: STUDY DESIGN, METRICS, AND REPORTING STANDARDS

CHAPTER 1

FEEDBACK: FEEDBACK LITERACY AND HUMAN-AI COLLABORATION

Introduction

Feedback is widely treated as one of the most powerful influences on student learning, but higher education has long struggled to make feedback timely, specific, dialogic, and usable at scale. Large classes, modular course structures, limited instructor time, and increasing assessment loads often produce delayed or generic feedback. Students may receive comments after the moment for action has passed, or they may be uncertain how to interpret feedback in relation to criteria, disciplinary expectations, and future tasks. Foundational feedback research has therefore moved beyond the idea of feedback as information transmission. Effective feedback is better understood as a process in which learners compare performance with standards, interpret information, regulate learning, and act on the feedback through revision or transfer (Butler & Winne, 1995; Hattie & Timperley, 2007; Shute, 2008).

Generative artificial intelligence (GenAI), especially large language models (LLMs), changes the practical conditions under which feedback can be produced. LLMs can generate rapid comments on drafts, explain rubric criteria, suggest revisions, summarize patterns in student responses, and support automated or semi-automated scoring of open-ended work. A recent systematic review suggests that AI-assisted feedback can address multiple feedback foci, including task, process, self-regulation, and self-level comments, and that the field has expanded sharply with the rise of modern LLMs (Ba et al., 2025). At the same time, empirical studies indicate that students and teachers do not experience AI feedback as a simple substitute for teacher feedback. Students often value AI feedback for accessibility, speed, volume, and lower social risk, but they also report concerns about reliability, contextual understanding, and trust (Henderson et al., 2025).

This tension is the starting point for the chapter. AI feedback is not automatically good feedback. A system that generates more comments may still fail if students do not understand, evaluate, or use those comments. A model that produces plausible scoring rationales may still be unreliable for higher-order judgment. A