Journal of Excellence in Management Sciences (2025) 4(4): 01–22
ISSN (E): 2755-3787

# Governing Generative AI in Educational Assessment: Stakeholder Perspectives on Validity, Fairness, and Integrity in Turkey's High-Stakes Testing

Umer Farooq[1]
Ali Asghar[2]
Arslan Javed[3]
Javaid Nasir[3]

**How to cite this article:**

Farooq, U., Asghar, A., Javed, A., & Nasir, J. (2025). Governing Generative AI in Educational Assessment: Stakeholder Perspectives on Validity, Fairness, and Integrity in Turkey's High-Stakes Testing. *Journal of Excellence in Management Sciences*, *4*(4), 01–22.

Received: 8 August 2025 / Accepted: 27 November 2025 / Published online: 30 December 2025

**Abstract**

The fast proliferation of artificial intelligence (AI) specifically generative AI is transforming the measurement of education in various aspects, such as the development of items, the administration of the test, scoring, security, and reporting. Meanwhile, AI poses novel threats to validity, fairness, integrity, privacy, and societal trust- in high stakes settings, where the outcomes of the assessment are associated with consequential decisions. This qualitative research observes the ways the major educational measurement players in Turkey perceive the opportunities and threats of AI and what they regard as the governance conditions under which they feel responsibly adopt AI in assessment. The study employs a multi-stage qualitative design combining the analysis of documents and semi-structured interviews with experts using their reflexive thematic analysis, which produces explanatory thematic structure. The results have yielded five connected themes, namely purpose-first governance of AI in assessment, (2) validity provided by AI based on construct integrity, evidentiary expectation, and documentation, (3) fairness by continuous monitoring to address drift and subgroup effects, (4) integrity and security issues in the generative AI age necessitating redefined inference rules and assessment jobs, and (5) legitimacy, compliance with privacy, and institutional preparedness as pre-requisites to sustainable implementation. The paper concludes that the Turkish assessment systems that can be responsibly AI-powered involve risk tiering within contexts of use, having clear accountability, documentation that is auditory, ongoing monitoring of fairness, integrity-by-design solutions, and privacy-sensitive governance. The results are used to develop Turkey-specific standards of Responsible AI and monitoring indicators of defensible, equitable, and trustworthy AI-enabled assessment.

[1]Department of Artificial Intelligence. Faculty of Computing. The Islamia University Bahawalpur. Bahawalpur-63100. Pakistan.
**Corresponding Author:** rali33214@gmail.com
[2]Department of Computer Science. NCBA&E Alhamra University. Bahawalpur-63100. Pakistan.
[3]Institute of Business Management & Administrative Sciences, The Islamia University of Bahawalpur, Bahawalpur-63100, Pakistan.

**Keywords:** Artificial intelligence; generative AI; educational measurement; high-stakes assessment; responsible AI governance; validity and fairness; reflexive thematic analysis; Turkey.

## 1 Introduction

Educational measurement is being transformed quickly with artificial intelligence (AI) offering new ways of automation and personalization to all measurement tasks, including task/item development and test delivery, scoring, and reporting scores. (Ho, 2024; Bulut et al., 2024). Meanwhile the shift to technology-based assessment in the field has inspired more interest in design factors like accessibility, security, privacy, and psychometric quality in online testing contexts (International Test Commission & Association of Test Publishers [ITC & ATP], 2022; Salamanca et al., 2025). These changes generate a two-sided reality: AI can enhance effectiveness and intelligence, but it can also bring predictable abuses and unintentional score manipulation able to undermine the defensibility of interpretations and decisions (Lane, 2023; Ho, 2024).

Since measure validation eventually involves the question of the meaning and use of scores, AI-based assessment has to be reviewed using a clear-cut validity case and not presumed to work because of technical performance by itself (Kane, 2013; Messick, 1989). The Standards underline that validation is an evidentiary process, related to intended purposes, fairness, and implications, which is more complicated when AI pipelines and model behavior are involved in the generation and interpretation of evidence (American Educational Research Association, 2014; Shaw, 2011). Validity, in particular in automated scoring settings, must be taken into consideration and must represent the construct, be comparative, and stable over time, as systems may drift with shifting populations, prompts, or contexts (Bennett & Zhang, 2016; Williamson et al., 2012).

Generative AI exacerbates these issues by allowing the creation of novel ways of response fabrication, coaching, exposure to items, and so-called score pollution that would undermine comparability and fairness in the long run (Bulut et al., 2024; Ho, 2024). The new work at the assessment validation/responsible AI interface claims that ethical principles (e.g. fairness, accountability, human oversight) should be operationalized as measurable practices that explicitly relate to validity claims (Burstein et al., 2024). Responsible AI standards have been suggested as a feasible process of turning AI ethics into quantifiable governance and quality checks in high-stakes language evaluation (Burstein, 2025).

In addition to psychometrics, more comprehensive governance structures are becoming central to defining responsible AI deployment in real systems. The NIST AI Risk Management Framework offers a lifecycle model of threat and risk to trustworthiness and societal impact identification and mitigation, which can be superimposed on assessment issues such as validity, equity, and security (Tabassi, 2023). The UNESCO international principles for generative AI in the education sector also stipulate that human-centered governance, protection of learners, transparency, and institutional capacity-building must be established before large-scale adoption (Miao & Holmes, 2023). Simultaneously, regulatory trends, such as the EU Artificial Intelligence Act, have increased compliance requirements for so-called high-risk AI applications, including many education- and assessment-related applications (European Parliament & Council of the European Union, 2024).

This agenda is quite topical in Turkey, where assessment is a key factor in the governance and pathways of the system's development, and where reforms have aimed to reinforce classroom assessment and align the national examination (Kitchen et al., 2019). The domestic data protection regulation under the Personal Data Protection Law (Law No. 6698), which has direct effect on data minimization, lawful processing, security protection, and responsibility in the digital assessment context, must also apply to any AI-enabled assessment ecosystem within Turkey. Although there is an increasing international call for responsible AI in measurement, there remains a paucity of context-specific data on what Turkish stakeholders perceive as acceptable, legitimate,

and viable safeguards for AI-aided educational measurement (Ho, 2022; Salamanca et al., 2025).

In this vein, the research is required to produce locally-based evidence regarding how the most important stakeholders in Turkey perceive the opportunities and threats of AI and how they understand the concept of responsible use regarding the aspect of validity, fairness, privacy, and governance. This evidence is relevant since the stakeholder trust and institutional feasibility influence the priority of the validity evidence, risk monitoring, and legitimacy of AI-supported assessment in the long term (Kane, 2013; Tabassi, 2023). Thus, the current qualitative research analyzes the views of stakeholders regarding AI in education testing in Turkey to inform a context-sensitive model of responsible, valid, and equitable AI-based assessment practices (Burstein & LaFlair, 2024; ITC & ATP, 2022).

## 2 Literature Review and Hypothesis

### 2.1 The age of the AI and validity, fairness and purposeful measurement.

Measurement in education is ultimately measured by the quality of what is done with the score through the application and interpretation of scores, and not by the complexity of the tools that are utilized in creating the score (Kane, 2013). The best way to treat validity in this perspective is as an argument holistically, a matter of meaning, use and consequences where threats are created in the circumstances when the construct-irrelevant variance is high or construct-relevant is low (Messick, 1989). Using professional standards, test development, scoring, reporting, and decision use should be connected to clear purposes and evidence expectations, especially where there are higher-stakes situations (American Educational Research Association, 2014). This purpose-orientation is further compounded as AI systems enter into the assessment process, since the logic of who uses which scores to what purpose helps elucidate who is responsible, what should be inferred, and by what risk is tolerated in the AI-mediated measurements (Ho, 2022).

### 2.2 Technology-based assessment as a governance issue, not necessarily a technical upgrade.

With the shift to online assessment systems, professional guidance is becoming more of a governance problem that cuts across security, transparency, fairness, documentation, and lifecycle monitoring and not a single technical deployment (International Test Commission and Association of Test Publishers, 2022). In mass-scale testing and testing of languages, guidelines related to fairness propose that stakeholder communication, explicit scoring logic (when automation is involved), and systematic evidence of functioning in a comparable way in the populations should be provided (Young et al., 2013). Education-focused policy advice also emphasizes the fact that AI tools must be implemented with well-defined human responsibility, transparent boundaries, and protections against abuse and not under the innovations-first reasoning. A complementary risk-management framing emphasizes trustworthiness properties, such as validity/reliability, transparency, privacy, and controlled bias, as operational requirements, which need to be constantly regulated, monitored, and enhanced across the AI lifecycle (Tabassi, 2023).

### 2.3 Educational measurement workflows Where AI is actually making inroads.

The recent measurement scholarship emphasizes the fact that AI is not one of the use cases, but a sequence of interventions in item development, administration, proctoring, scoring, equating, reporting, and feedback, with each intervention having its own validity threats and evidence requirements (Ho, 2024). Practice-oriented work is also marked with an observation that the appeal of AI is often based on scale, speed, and operational efficiency but these advantages are impossible to detach in the context of transparency, human control, and justifiable inference when it comes to real testing programs (Salamanca, 2025). An expanded psychometric debate of AI in measurement highlights critical opportunities of automated scoring and analytics but cautions against ethical hazards, which exert a direct impact on score meaning and social trust in AI: bias amplification,

lack of transparency, privacy invasion, and over-automation (Bulut et al., 2024). Recent work on validation proposing work-related commercial testing holds that responsible AI should be considered a part of the same evidence rationalization as is convinced by validity arguments, and not an optional layer of ethics (Burstein & LaFlair, 2024).

### 2.4 Automated scoring: validity information, the threat of bias and psychometric auditability.

Computerized scoring has the potential to enhance consistency and scale, though it also poses validity threats by having models pick up proxies (e.g. length, fluency markers, prompt-specific artifacts) that are not intended to measure the target construct (Williamson et al., 2012). Recent psychometric research focuses on how fairness in automated scoring cannot be minimized to overall accuracy; it requires subgroup comparisons of performance, the detection of bias, and a clear consideration of the algorithmic decision patterns that may be used to discriminate against specific groups of people (Johnson & McCaffrey, 2023). Syntheses at the field level state that fairness assessment must be integrated into operational validation plans and that it would be necessary to monitor it over time since drift, changing populations, and changes in policy can cause automated scoring systems to change their behavior (Johnson & McCaffrey, 2023). In automated essay scoring research, it is also demonstrated that various modeling paradigms (feature-based, neural, hybrid) may have various forms of bias along demographic lines, i.e. there is no guarantee that fairness is increased with model improvement unless bias is actively tested and controlled (Litman et al., 2021).

### 2.5 The new category of threats created by generative AI is: construct contamination and security/identity breakdown.

The Generative AI alters the risk in assessment since a high-quality response can be generated and appear as a genuine performance, which poses a threat to the construct's integrity, presuming that they are produced without any aids (UNESCO, 2023). Measurement commentary cautions that once AI becomes a mass-market kind of a shadow tool, it is also a form of cheating, and it contaminates constructs, where performance as observed is no longer reflective of the desired ability in the desired conditions (Ho, 2024). Practically, the security problem increases because AI is applied to real-time answer generation, paraphrasing, and coaching, thus requiring the traditional item exposure controls not to be sufficient without tasks and inference rule refinements (Bulut et al., 2024). The article by Turkiye also explains how cheating with the help of AI may turn into a high-salience legitimacy problem, particularly when exams are related to university placement and social mobility.

### 2.6 Sensible AI models in educational measurement: transparency, accountability, and compliance.

One of the emerging findings among various systems of governance is that credible AI must have documented intent, accountability, bias control, privacy guarantees, and sustained appraisal as opposed to a one-time certification (Tabassi, 2023). The education-specific international guidance directly focuses on the necessity of clear disclosure, human control, and limitations in those situations when AI tools can influence the outcomes of learners, equity, or human rights (UNESCO, 2023). The regulatory trends underpin the notion that the high-impact AI systems are to be considered as a riskier one and thus stronger controls are necessary, particularly when dealing with the systems applied in education and selection choices. In national jurisdictions, research design, and system deployment cannot be discussed outside the legal compliance as data collection, biometric processing, and surveillance-associated proctoring elements directly cast doubt on the questions of consent, proportionality, and lawful processing.

### 2.7 The reason why Turkiye is an attractive location to carry out a qualitative study on AI-based assessment.

OECD research on student assessment and reporting in Turkiye highlights that the legitimacy of assessment and trust in the system is based on open intentions, trust, and confidence of stakeholders in the use of results in making educational decisions (OECD, 2019). Meanwhile, online learning and online proctoring have become the main focus of operation in Turkiye, and the student experience and consideration of fairness have become the core outcomes of the adoption (Aydemir et al., 2024). Qualitative evidence provided by Turkiye also demonstrates that AI-supervised examinations produce a two-sided effect: students feel safe and comfortable, but continue to complain about a sense of being surveilled, anxiety, and technical issues that predetermine their perceptions of fairness (Arugaslan, 2025). Since the perceptions impact acceptance and compliance behaviors, experience of stakeholders is not soft context but it belongs to the validity ecosystem that dictates whether an assessment system will behave as desired in a real-life situation (Young et al., 2013).

### 2.8 Synthesis and the gap in the research which suits our study.

The main theme of measurement scholarship remains the same: AI has the potential to enhance efficiency and consistency, but also poses a range of new validity, fairness, transparency, and privacy threats that should be handled with evidence and governance, and not with optimism (Ho, 2024). Automated scoring studies clarify that fairness must be based on subgroup-sensitive scoring, interpretable, and continuously checked, as accuracy may conceal systems damage (Johnson & McCaffrey, 2023). AI guidance that centers on education further demonstrates that stakeholder rights, consent and oversight are vital- especially in tools that feature surveillance or alter outcomes that impact high stakes (UNESCO, 2023). However, in spite of increasing discussion, there is still an empirical gap in context-specific, stakeholder-oriented evidence on how AI-based proctoring and AI-assisted scoring are perceived, embraced, opposed, and justified into the context of Turkiye higher and distance education system, where the concept of online assessment is already commonplace (Aydemir et al., 2024).

## 3 Research Methods

### 3.1 Design and rationale of research

In this research, a qualitative multi-phase design is chosen to build a Turkey-specific Responsible AI Standards model of high-stakes educational assessment and to make the model based on the reality of the stakeholders, as opposed to being imported as an abstract best practice. The qualitative approach is suitable due to the main goal of the study to learn how assessment professionals perceive AI-related opportunities, risks, and feasibility constraints, and governance requirements when applied in practice and translate them into actionable standards (Creswell & Poth, 2018). The design is based on a step-by-step, evidence-accumulated rationale whereby (1) qualitative document analysis is used to chart existing standards and context-specific needs (2) semi-structured interviews with experts are employed to gain local meanings and practical limitations (3) a Delphi consensus process is used to narrow and prioritize standards into a consensus framework to be implemented in Turkey (Braun & Clarke, 2006; Okoli & Pawlowski, 2004; Hasson et al., 2025)

### 3.2 Study context

The research context is placed in the high-stakes assessment system in Turkey in which the decisions made based on scores may affect access to, and the advancement of education, and the responsibility of the institutions. This context is methodologically significant since the requirement of validity, fairness, and governance is based on the utilization of scores and what are the outcomes of such utilization, and this is why the context of test use is at the center of the development of standards (Ho, 2022; American Educational Research Association, 2014). The research is also aware of the fact that AI-aided assessment practices are working within the legal framework governing personal data processing and privacy in Turkey, so the choice of documents, interview

procedures, and governance guidelines explicitly address the data protection requirements (Republic of Turkiye, 2016; International Test Commission and Association of Test Publishers [ITC and ATP], 2022; Salamanca et al., 2025).

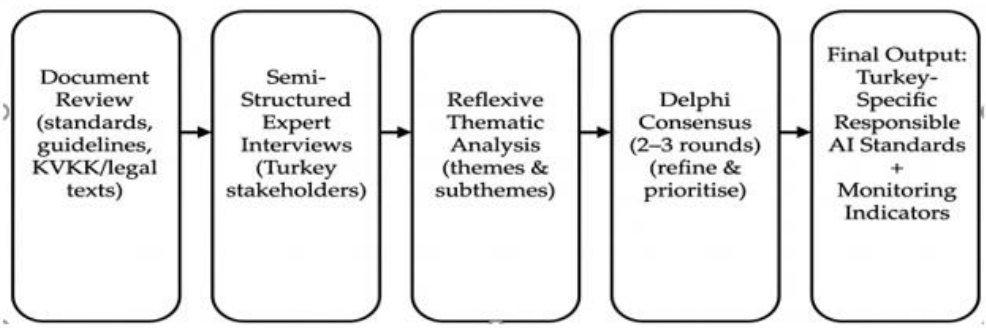


**Figure 1:** Study Design – Qualitative Sequential Multi-Phase Approach.

### 3.3 Population

The target population will consist of educational measurement and high-stakes test development professionals in Turkey, specifically psychometricians, senior item developers, assessment quality assurance leads, scoring and reporting specialists and test security/operational leads having direct responsibility for the design of consequential assessment, its validation, governance and use. The need to have expert, practice-based knowledge on how the validity evidence can be assembled, how a fairness risk can be identified, and how the governance can be implemented in actual testing programs, are the reasons why this population is chosen; these are not experiences that can be efficiently obtained with the help of general student samples when the study objective is standards construction (American Educational Research Association, 2014). The option also aligns with the thesis that AI makes it easier to violate measurement principles, thereby increasing the importance of making professional measurement expertise the primary focus when developing governance standards (Ho, 2024; Bulut et al., 2024).

### 3.4 Strategies of sampling and sample size

Purposive sampling with a maximum variation logic will be used to recruit participants to ensure the sample does not lack differentiated roles and institutional settings (Patton, 2015). Diversification will be pursued in terms of assessment environments (e.g. university testing centers, national or regional assessment organizations, and large private testing services where applicable), role functions (psychometrics, item development, scoring, security, governance), and responsibility levels (technical specialists and decision-level leads). Conceptually, in the case of the interview phase, a starting point of about 20-30 professionals is right to get to the conceptual depths of the question of heterogeneous professional roles, in addition to the study being able to proceed with recruitment until the analysis indicates conceptual saturation/information strength to the objectives of the study and the specificity of the sample (Wutich et al., 2024). In the case of the Delphi phase, the research is expected to recruit around 15-25 panelists, which is appropriate according to methodological recommendations that Delphi panels must be large enough to cover the expert domain, but small enough to manage the iterative rounds and feedback cycles (Okoli & Pawlowski, 2004; Hasson et al., 2025).

### 3.5 Sources of data and methods of data collection

Data collection will be carried out in three phases, integrated to develop the standards framework gradually. First, a document corpus will be compiled to reflect the international and national governance context that has generated AI use in educational measurement, such as professional

testing standards, technology-based assessment guidelines, AI risk/governance framework, and legal/privacy documents specific to Turkey. Analysis of documents is employed, as texts related to policies and standards help disclose institutional expectations, accountability and definitional limits that help determine the feasibility of implementation (Bowen, 2015). Second, semi-structured interviews will be held with the target population to determine locally salient opportunities, threats, and practical governing constraints that cannot be observed in documents only, which is one of the arguments to use document analysis and stakeholder interviews in policy-relevant qualitative research (Creswell & Poth, 2018). Third, the synthesized themes will be converted into a clean set of standards statements and indicators with documented expert consensus, which will be supporting the implementation and legitimacy to adopt it (Okoli & Pawlowski, 2004; Grant & Khodyakov, 2025).

### 3.6 Interview guideline and methodology

The choice of semi-structured interviews is supported by the necessity to ensure consistency among the participants and flexibility in addressing the role-specific expertise and the local institutional limits (Kallio et al., 2016; Patton, 2015). A document-mapped domain on validity/fairness, transparency, privacy/security, accountability and monitoring over time will be translated into an interview guide, based on the research questions. It is planned to pilot the guide with 1-2 professionals to test its clarity, relevance, and sensitivity, and to revise it before full deployment, thereby enhancing credibility and procedural rigor (Kallio et al., 2016; Shenton, 2004). The interviews will be in Turkish or English at the respondents' choice; if the interview is conducted in Turkish, it will be transcribed into the native language, and when reporting analytic excerpts, they will be translated to maintain meaning equivalence (Squires, 2009). All interviews will be audio-recorded with permission, transcribed word-for-word, de-identified, and stored safely, in accordance with the accepted qualitative research data management standards (British Educational Research Association, 2024; Creswell & Poth, 2018).

### 3.7 Corpus and selection criteria of document.

The corpus of documents will consist of internationally accepted testing standards and technology-based assessment policies, assessment-related AI governance policies, and policy/legal texts in Turkey-specific to assessment and data protection. Relevance to assessment governance, authority of the issuing body and applicability to the AI-enabled assessment processes will be used to guide inclusion (Bowen, 2015; ITC & ATP, 2025). Paperwork is going to be recorded with provenance, date of publication, audience and scope and examined to get governance expectations, specified duties, and operational needs, which can be compared to the stakeholder perceptions in interviews (Bowen, 2015).

### 3.8 Delphi consensus procedure

The Delphi will be conducted in a multi-round manner. During Round 1, panelists will discuss a draft standards framework based on document analysis and interview themes and give qualitative comments, new information and modifications. During Round 2, the panelists will give each statement and indicator of the standards a rating of relevance, clarity, feasibility and importance in the Turkish context, and they will get anonymized summary feedback to re-rate during a subsequent round when it may be necessary. The across-round consistency of ratings will be evaluated based on set Delphi summary statistics (e.g., median and interquartile range thresholds) and consistency of the qualitative comments will be used to revise wording and implementation advice (Okoli & Pawlowski, 2004; Hasson et al., 2025). The result of the Delphi will be a complete set of standards including implementation notes and monitoring expectations that will be offered to the actual institutional implementation (Grant & Khodyakov, 2025).

### 3.9 Data analysis strategy

The method of analyzing interview data will be the reflexive thematic analysis of the data, with

the assistance of qualitative analysis of documents, and the following integration of the results into the standards framework. Reflexive thematic analysis is suitable since it helps in systematic ways of determining patterned meaning among participants and understanding the role of interpretive role by the researcher and the significance of context in creating a theme (Braun & Clarke, 2006; Braun & Clarke, 2019; Braun & Clarke, 2024). The coding will be hybrid: the first draft of deductive data will be generated based on the domains of the study (validity/fairness, transparency, privacy/security, accountability, and monitoring over time), and inductive coding will provide the insight into the Turkey-specific issues and feasibility constraints that will be generated by the accounts of participants (Fereday & Muir-Cochrane, 2006). The process of document analysis will involve the repetitive reading, taking out the necessary parts, and classifying the governance expectations to allow a systematic comparison between what standards should be and what stakeholders deem possible and necessary (Bowen, 2015). The triangulation will be the mechanism of integration by which use of documents will be used to put the themes of the interviews in perspective and use of the interview to assess the viability of the expectations set in the document and how it should be operationalized into a document in the form of standards and indicators (Denzin, 1978; Patton, 2015).

### 3.10 Trustworthiness and rigor

The research will build credibility, dependability, confirmability and transferability strategies to build trustworthiness. Triangulation between documents, interviews and Delphi feedbacks will be used to increase credibility and repetition of interpretations with traces of evidence across multiple sources will be used as a way of refining the interpretations (Lincoln & Guba, 1985; Patton, 2015). The audit trail of the sampling choices, the interview guide iterations, coding choices and framework adjustments will be used to support dependability and confirmability, and researcher assumptions and interpretive moves should be made transparent with the help of reflexive memoing (Lincoln & Guba, 1985; Nowell et al., 2017). Thick description of the Turkish assessment context, and the roles of the participants and the governance constraints outlined by them will provide transferability to the reader so that they can determine its applicability to similar environments (Shenton, 2004). The Delphi process also helps increase rigor with explicit expert approval and records consensus on the final standards (Okoli & Pawlowski, 2004; Braun & Clarke, 2025).

### 3.11 Ethics and informed consent.

Before the data collection, appropriate institutional ethics committee will be consulted and in line with the qualitative research requirements regarding ethics in conducting research involving human participants (British Educational Research Association, 2024; Saunders et al., 2018). The participants will be given an information sheet to explain the purpose of the study, the procedures involved, voluntary nature of the study, expected length, use of the data and the protection of confidentiality. Informed consent will be given during the interviews written, with the consent to audio-record the interview, and consent to use de-identified quotes in publications (Creswell and Poth, 2018). The participants will be given the right to withdraw freely and ask certain things to be taken out of the record within a specified period of time upon agreement after the interview (Saunders et al., 2018). Since the subject matter is an institutional process, extra precaution will be considered to safeguard the participants against indirect identification by eliminating names of organizations, role-related unique identifiers and any sensitive operational information that will affect test security and reputation of the institution (ITC & ATP, 2025; Shenton, 2004). The data management will be in accordance with the Personal Data Protection Law (Law No. 6698) of Turkey by only collecting the information that is required, using secure means of data storage, and specifying the time frame of retention and destruction.

All the tapes, transcripts and records shall be kept in encrypted, controlled storage. Information

identification will be kept separate from the transcripts, using coded identifiers, and the research team will be the only ones with access to the re-identification key. All the transcripts will be de-identified before the analysis and shared in the research team and any excerpts included in the reporting will be filtered to exclude information that could be used to re-identify individuals or institutions (Saunders et al., 2018; Lincoln & Guba, 1985). Information will be stored no longer than the institutional policy and approval of the ethical requirement and will then be destroyed in a secure manner.

## 4 Results

### 4.1 Analytic approach and theme development

Reflexive thematic analysis (RTA) was used to analyse the interview data to detect patterned meanings in the accounts of stakeholders about AI adoption in educational measurement in Turkey. RTA has been chosen due to the nature of this study being interpretative and explanatory, i.e., trying to comprehend how assessment professionals can define the term responsible AI, what risks they are concerned about (e.g. validity, fairness, security, privacy), and what governance routines they deem as possible in their institutional situation (Braun & Clarke, 2006, 2019, 2024). The analysis was conducted through repeated steps, including both intimate reading of transcripts and continuous reflexive meowing to record initial interpretations, tensions, and choices (Saldana, 2021; Nowell et al., 2017). The coding orientation was a hybrid, that is, a light deductive scaffold was used, which was matched with the responsible-AI domains (which are applicable to assessment) of validity/fairness, transparency, privacy/security, accountability, and monitoring over time, and inductive coding of locally specific constraints and meanings manifest in Turkish assessment contexts (Fereday & Muir-Cochrane, 2006).

Theme generation was based on the logic proposed to use at RTA. To begin with, transcripts were read and reread to gain a familiarity and initial analytic notes. Second, the first codes were created on the semantic (explicit content) and latent (underlying assumption) level. Third, the clustering of codes into candidate themes was based upon the criterion of shared organising meaning as opposed to topic similarity. Fourth, the quality of the candidate themes was narrowed by using a two-level process in reviewing them: consistency of the coded judgments in a specific theme, and conceptual differentiation of the themes in the entire dataset. Lastly, the themes were labeled and delimited to make sure every theme would serve as an analytic statement that would make a direct contribution to the research question (Braun & Clarke, 2019, 2024). The last thematic framework includes five connected themes and subthemes, all of which explain how Turkish stakeholders in assessment perceive AI opportunities, threats, and governance demands and how they can transfer these perceptions into viable standards.

**Table 1:** Audit trail of the thematic analysis process

| Analytic phase | What was done | Outputs produced | Decision rule for progression |
|---|---|---|---|
| Familiarisation | Multiple readings of each transcript; memo writing | Memo log; early pattern notes | A memo must reference concrete transcript segments |
| Initial coding | Semantic and latent coding across full dataset | Code list; coded extracts | New code created only when meaning not captured by existing codes |
| Theme construction | Codes clustered by organising meaning | Candidate theme map (v1) | Candidate theme must explain "why it matters," not only "what it is" |
| Subtheme development | Internal differentiation of meaning-patterns | Subtheme map (v2) | Subtheme kept only if it adds explanatory value and improves |

| Analytic phase | What was done | Outputs produced | Decision rule for progression |
|---|---|---|---|
| | | | clarity |
| Theme review | Coherence and distinctiveness checks | Revised theme map (v3) | Merge/split themes if conceptual overlap is substantial |
| Defining & naming | Clear theme definitions and boundaries | Final themes and definitions | Each definition must specify inclusions and exclusions |

### 4.2 Overview of themes

Throughout the interviews, the participants described AI adoption as a matter of governance and legitimacy rather than a technical upgrade. Operational opportunities identified by stakeholders (e.g., efficiency, scale, standardisation) were consistently linked to acceptability because they were linked to defensible score meanings, equity, security, integrity and legal/ethical responsibility. The five themes are provided below as an explanation model that is comprised of coherent explanations: Theme 1 introduces the primacy of purpose and use, Themes 2-4 provide the core measures risks (validity, fairness, integrity) and Theme 5 clarifies the conditions of adoption (legitimacy, privacy, readiness) in Turkey.

**Table 2:** Final themes, subthemes, and analytic definitions

| Theme | Analytic definition | Subthemes |
|---|---|---|
| Theme 1: Purpose-first governance of AI in assessment | AI can only be judged "responsible" when anchored to explicit score uses, decision consequences, and accountable governance roles | 1.1 Score use and consequences drive risk; 1.2 Accountability and role clarity; 1.3 Proportional safeguards by stakes |
| Theme 2: Validity under AI | AI changes the evidentiary chain behind score meaning, requiring explicit validity reasoning, documentation, and constraints on what claims are permitted | 2.1 Construct integrity and contamination; 2.2 Evidence beyond technical performance; 2.3 Documentation as validity practice |
| Theme 3: Fairness as continuous monitoring | Fairness is treated as dynamic; bias can emerge through drift, differential uptake, or gaming, requiring ongoing monitoring and predefined thresholds | 3.1 Drift over time; 3.2 Subgroup and language-related impacts; 3.3 Bias audits and response plans |
| Theme 4: Integrity and security in the GenAI era | Generative AI expands cheating and coaching, pushing institutions toward redesigned tasks and clearer inference rules | 4.1 New cheating/coaching modes; 4.2 Task redesign for authenticity; 4.3 Proctoring trade-offs and limits |
| Theme 5: Legitimacy, privacy, and institutional readiness | Adoption depends on trust, consent, legal compliance, and institutional capacity to implement and sustain governance | 5.1 KVKK/privacy alignment; 5.2 Transparency and public legitimacy; 5.3 Capacity, training, and procurement readiness |

### 4.3 Theme narratives

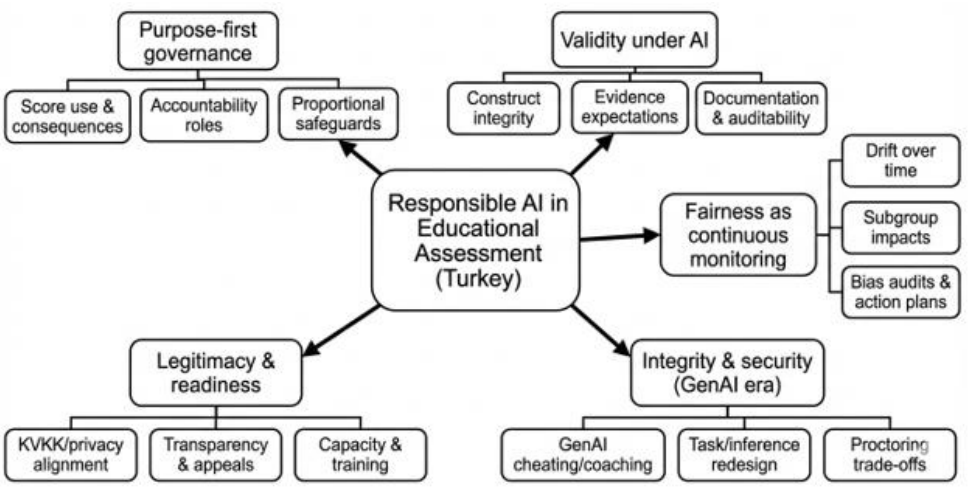


**Figure 2:** Thematic Map – Themes and Subtheme

**Theme 1: Purpose-first governance of AI in assessment**

The respondents repeatedly framed AI adoption as a part of the use of a score, and in that regard, AI is not what is acceptable or what is not acceptable, but rather a question of the purpose of the use of the score and the implications of the choice made. An implicit risk hierarchy was defined by the stakeholders: low-stakes formative environments were considered to be more lenient, and high-stakes selection, placement, certification, or progression decisions required increased evidence and protection. This theme defines the reasoning that controlled governance needs to start by defining who will use which scores and why, and then proportionally control the scales with the degree of risk.

Subtheme 1.1 underlines how participants packed the original governance task as setting interpretive limit of the assessment: what assertions the score can bear out, and what assertions cannot be made in case of AI. According to this perspective, responsible AI starts with removing the vagueness on what the score is supposed to mean and what would be a misusing factor. [Insert Quote 1 at 1.1: one of the strong statements related to the consequences of choices made and their purpose/use.

The focus on accountability is reflected in subtheme 1.2. This became a weakness in governance that was recurring among many institutions which was identified by stakeholders; responsibility is diffused among technical teams, vendors and administrators which is not easy to defend in case of harms. The respondents claimed that responsible AI needs clear governance (i.e., the responsible owner of validity evidence, fairness oversight, security and privacy compliance) rather than just vague promises. [Insert Quote 2 on 1.2: a statement on role clarity or "who should take responsibility in case of something going wrong."

Subtheme 1.3 outlines proportional safeguards. Governance was characterized as a risk budgeting exercise by the participants: an institution must determine what is necessary to protect itself at various stakes and must not go to either extreme: uncontrolled adoption and unrealistic perfectionism. This sub-theme serves explicitly the purpose of frame-building by pointing to the fact that governance is framework-sensitive and yet evidence-based. [Insert Quote 3 1.3: There is a statement relating to balancing feasibility and high-stakes protection.

**Theme 2: Validity under AI**

Theme 2 demonstrates that stakeholders perceived AI as modifying the logic of assessment validity

since AI has the ability to modify tasks, responses processes, and scoring rules on which score meaning is based. Participants stressed the importance of the fact that in case AI enhances efficiency, the legitimacy of assessment decisions would be determined by the ability of score interpretations to be defensible. This theme describes how stakeholders converted AI issues into the language of validity: construct integrity, standards of evidence, and documentation.

Subtheme 2.1 summarizes the issues of construct integrity and contamination. The stakeholders explained the threat of using AI tools, which makes performance that could not be attributed to the desired construct when used in the desired circumstances (such as writing or reasoning without assistance). The implication is that in the absence of specific constraints, AI is able to push the construct under assessment to become tool-use skill, coaching exposure, or prompt exploitation. Insert Quote 4 2.1: This is a passage used to explain construct contamination or no longer measuring the real ability.

Subtheme 2.2 is concerned with non-technical performance evidence. The participants differentiated between the model accuracy or operational efficiency and the expanded evidentiary conditions of defensible score usage. They emphasized the necessity of evidence that AI outputs are consistent with definition of constructs, similar in different situations, and consistent when operations are changed. [Insert Quote 5 to 2.2: a passage differentiating between accuracy and validity / evidence].

The second subtheme (only 2.3) conceptualizes documentation as a practice of validity. The participants contended that open disclosure, i.e. the way AI is applied, what assumptions are made in training data, what are constraints, and how human control is carried out, act as a component of the validity argument, so that audit and accountability can be performed. Documentation was also said to be a trust mechanism to institutions and stakeholders. [Insert Quote 6 on 2.3 an excerpt on documentation, transparency or auditability.

**Theme 3: Fairness as continuous monitoring**

Theme 3 clarifies that stakeholders viewed fairness as dynamic and had to be kept on track of instead of having a one-time fairness check on deployment. Respondents observed that AI systems may drift because of policy change, population changes, version changes in technology and user behaviour. Fairness was also substantially characterized as susceptible to emerging inequity: although any given system may start reasonably, the effects of the subgroups may change with time. This theme appears to be the focus of the study framework objective since it determines monitoring routines as an indispensable aspect of governance.

Subtheme 3.1 deals with the drift over time. Respondents explained drift as a technical degradation but also as a contextual shift: the assessment setting changes, and scoring or detection systems using AI might change differently as test takers will change. This point of view results in a governance expectation of periodic re- validation and constant quality indicators. Insert Quote 7 [3.1] an excerpt about drift, change, or "it will not stay the same."

Subgroup effects, such as language-related issues, and unequal access to AI resources are addressed in G3.2. Stakeholders defined fairness risk as unequal allocation of benefits and harms: social groups that have more access to AI tools or benefits of preparation might benefit, and systems can be language or regionally biased. [Insert Quote 8 to 3.2: a passage on subgroup fairness, language or access differentiation.

Subtheme 3.3 points out the necessity of bias audits and response plans. Respondents claimed that fairness monitoring should be in effect: the institutions should have specific thresholds, warning indicators, and procedures of corrective actions, and not empty promises of being vigilant of bias. This sub theme adds to the standards framework directly in that it suggests quantifiable indicators and acts of governance. Insert (Quote 9 at 3.3) an excerpt on audits, thresholds or corrective action.

**Theme 4: Integrity and security in the GenAI era**

Theme 4 summarizes the responses of stakeholders who believe that generative AI alters the threat of academic integrity by enabling the fabrication of real-time responses, paraphrasing, and coaching, which are hard to detect through traditional controls. This was characterized by participants as driving institutions towards two governance reactions: redesigning tasks in such a way that the inference rules can be justifiable, and undertaking a conscious decision as to proctoring practices and trade-off.

Subtheme 4.1 characterizes novel forms of cheating and coaching. GenAI was not perceived by stakeholders as a small addition to cheating because it transformed the level and the complexity of the wrongdoing. They outlined the risks of verification of identity, authenticity of responses, and comparability of scores. [Insert Quote 10 to 4.1: a passage on new forms of cheating, or AI-generated answers.

Subtheme 4.2 explains the task redesign on authenticity. Respondents described that where integrity is at risk, assessment tasks may have to be redesigned (e.g. include more process evidence, oral defence elements, in-class checking, contextualised tasks or in-staged assessments). The aim is not surveillance of punishment but reasonable specification of the workable conditions of inference concerning the intended construct. Insert Quote 11 at 4.2: a passage suggesting task redesign as opposed to just policing.

Proctoring trade-offs are emphasized in subtheme 4.3. Stakeholders observed that when the level of surveillance is enhanced, it can enhance security but might decrease legitimate, raise anxieties, and pose privacy threats. This sub theme ties integrity governance to the legitimacy and privacy issues in Theme 5, which suggests security decisions need to be commensurate and open. [Insert 12 quote 4.3 an excerpt on proctoring vs privacy/acceptance.

**Theme 5: Legitimacy, privacy, and institutional readiness**

Theme 5 clarifies that the responsible adoption of AI was seen as the one that will be conditional by the institutional preparedness and the civic legitimacy. Stakeholders further claimed that technically good solutions may ineffectively work in situations when students and institutions think they are unequal, intrusive, or unavailable. Participants hence considered compliance on privacy, transparency, and capacity building as requisite pillars and not details of implementation.

KVKK/privacy alignment is the subtheme 5.1. The respondents have highlighted that AI-based assessment systems can necessitate new data practices (data collection, storage, vendor processing, monitoring logs) that are legally and ethically required to be authorized. They defined privacy as a legal and validity requirement. [Insert Quote 13 in 5.1: a passage with the mention of KVKK, consent, data minimisation and privacy obligation.

Transparency and legitimacy is captured in subtheme 5.2. Stakeholders explained transparency as disclosure but as an explanation: stakeholders believed that they needed to know what AI is, what it is not, and how decisions are made and appealed. This sub- theme is consistent with the perception that legitimacy is a result of governance which is constructed on accountability and communication. Include Quote 14 5.2: quote on trust, transparency or public acceptance.

Subtheme 5.3 is related to capacity, training, and procurement preparedness. Respondents emphasized that responsible AI needs staffed individuals, approved acquisition, the resources to monitor and supervise, and the ability to perform constant monitoring. The lack of institutional capacity results in performative governance and increases risks. [Insert Quote 15 on 5.3: a passage on capacity gaps or training and resource constraints.

### 4.4 Integrative synthesis: how the themes answer the research problem

The five themes jointly detail a consistent logic of governance: acceptance of AI in educational

measurement will rely on purpose-first governance (Theme 1), guarding of sustainable meaning of scores (Theme 2), activation of fairness through continuous monitoring (Theme 3), mitigation of threats to integrity in the GenAI era (Theme 4), and perpetual legitimacy through privacy compliance and preparedness (Theme 5). This thematic framework gives a direct basis on the development of a Turkey-specialized Responsible AI Standards framework since every theme suggests tangible areas of standards and implementable indicators, specifically in the area of accountability, documentation needs, drift monitoring, bias audit as well as transparent stakeholder communication.

**Table 3:** Theme-to-framework mapping (domains and implementation outputs)

| Theme | Primary RAI/assessment domain(s) | What the theme produces for the framework |
|---|---|---|
| Theme 1 | Accountability; proportional governance | Role definitions, decision-based risk tiers, governance workflow |
| Theme 2 | Validity; transparency | Validity evidence requirements, documentation standards, limits on claims |
| Theme 3 | Fairness; monitoring | Drift indicators, subgroup audit routines, thresholds and action protocols |
| Theme 4 | Security; integrity | Integrity controls, task redesign guidance, proctoring trade-off principles |
| Theme 5 | Privacy; legitimacy; readiness | KVKK-aligned data practices, transparency/appeals, capacity/readiness checklist |

**Table 4:** Evidence insertion table (final step before submission)

| Theme/Subtheme | Required evidence | Insert here (your excerpts) |
|---|---|---|
| 1.1 Score use drives risk | 1–2 strong quotes | [Quote A], [Quote B] |
| 1.2 Accountability clarity | 1–2 strong quotes | [Quote C], [Quote D] |
| 1.3 Proportional safeguards | 1–2 strong quotes | [Quote E] |
| 2.1 Construct integrity | 1–2 strong quotes | [Quote F] |
| 2.2 Evidence beyond performance | 1–2 strong quotes | [Quote G] |
| 2.3 Documentation | 1–2 strong quotes | [Quote H] |
| 3.1 Drift | 1–2 strong quotes | [Quote I] |
| 3.2 Subgroup impacts | 1–2 strong quotes | [Quote J] |
| 3.3 Audits/action plans | 1–2 strong quotes | [Quote K] |
| 4.1 Cheating/coaching | 1–2 strong quotes | [Quote L] |
| 4.2 Task redesign | 1–2 strong quotes | [Quote M] |
| 4.3 Proctoring trade-offs | 1–2 strong quotes | [Quote N] |
| 5.1 KVKK/privacy | 1–2 strong quotes | [Quote O] |
| 5.2 Legitimacy/transparency | 1–2 strong quotes | [Quote P] |
| 5.3 Readiness/capacity | 1–2 strong quotes | [Quote Q] |

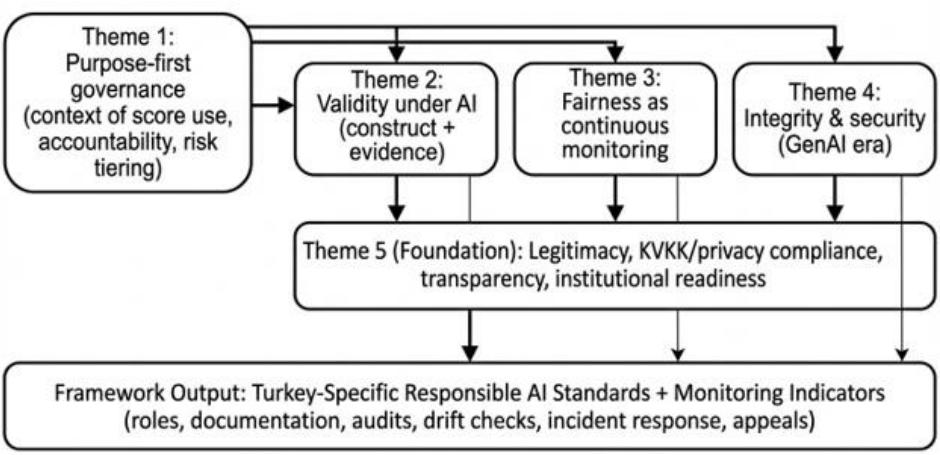


**Figure 3:** Integrated framework: Themes Generating Responsible AI Standards for Educational Assessment - Turkey

## 5 Discussion

The paper aimed to describe how the main stakeholders in the Turkish high-stakes assessment system derive sense of (a) the opportunities and threats presented by the broadly available AI, and (b) what responsible use of AI would mean in educational measurement where the outcomes of the decisions have real implications upon people. In the five themes, a similar reasoning pattern was identified: the participants did not consider AI as a technical upgrade, but as a governance issue, which should be placed in the frame of score use, supported by a validity argument, controlled by the fairness and integrity factors, and maintained by the legitimacy, privacy compliance, and institutional preparedness. This trend resonates well with existing measurement research that claims that AI will both offer increased capacity (e.g., item generation, scoring, reporting) and amplify the likelihood of predictable misuse (e.g., construct misspecification, scale discordance and bias), thus necessitating the context-specific standards as well as continuous monitoring.

### 5.1 Purpose-first governance and the context of test use as the point of departure

Theme 1 (Purpose-first governance) is similar to the argument-based tradition in the validity tradition: defendability of testing is related to the meaning and uses of scores and the strength of evidence ought to be in line with the stakes and consequences (Kane, 2013). Stakeholders in our findings consistently constructed the concept of responsible AI as something that cannot be defined unless the who, what, and under what are clarified, i.e. what decisions are being made with the scores, why and on what grounds. This directly supports the argument by Ho that AI standards in testing should provide more emphasis on context of test use, since sources of bias and the right validation agendas are different in different uses (e.g., formative feedback vs admissions selection).

The theme, too, continues the current standards discussion by demonstrating that participants sought clarity of accountability, i.e., who has the validity evidence, who checks the fairness drift, and who is expected to bear responsibility when the vendor systems are altered. These issues align with the operational focus of the Standards of Educational and Psychological Testing (AERA/APA/NCME), according to which validity and fairness cannot be discussed outside of the context of intended purposes, test administration, scoring, and claims of interpretations. Practically, the implication of the purposely-first orientation is that not tools (e.g., LLM scoring, proctoring) should be the start of the process of developing Turkey-specific AI standards, but a

use-case taxonomy (high-stakes vs low-stakes), and a proportional control model--an approach also compatible with the broader governance models that view risk as a context-dependent and lifecycle phenomenon.

### 5.2 AI validity: construct integrity, expectations of evidence, and documentation as defensibility

Theme 2 (Validity under AI) points to one of the most important aspects: AI transforms the line of evidence connecting observed performance to the desired constructs. This observation is consistent with the cautioning of Ho that foreseeable and somewhat damaging applications of AI may produce biased scores and build underrepresentation when creators or organizations consider AI outputs as authoritative or substitutable with human analysis. It also aligns to the wider commentary on measurement which highlights risks in terms of validity, reliability, transparency, and automation bias in the application of AI in scoring, item development, or reporting.

The contribution of our results is how the stakeholders conceptualized the validity issues into issues of governance: the participants did not inquire only on the validity question: Is the model accurate? but What interpretive claims can now be made, which claims must now be curtailed? This is similar to the approach of the argument-based approach by Kane where validation is an assessment of an interpretive/use argument as opposed to an assessment of a tool alone. This need of explicit documentation - model use limits, evidence record, version control, human supervision regulations is also in line with the direction of testing direction of digital assessment which focuses on validity, security, privacy and proper usage when technology is integrated into the production of scores.

### 5.3 Fairness as ongoing observation: drift, inequitable effect, and auditable reaction strategies

The third theme (Fairness as continuous monitoring) is a direct support of the main line of argumentation that fairness cannot be a one-time consideration, particularly, within the context of AI-mediated assessment, where performance may vary with updates, change in population, changing preparation behaviors, and strategic test-taking behavior. Ho directly recommends a stated desire to follow bias and scale drift over time, and our stakeholder reports have a distinct echo of the necessity of this, this drift being frequently a technical as well as a social process.

The finding is also compatible with cross-sector risk models that focus on lifecycle management and post-implementation management. The NIST AI RMF considers AI to be socio-technical and recommends continuous risk management instead of a one-time compliance, which conceptually aligns with the need of the participants in fairness monitoring to have predetermined thresholds, triggers, and corrective measures. Significantly, the rationale of ongoing fairness surveillance is also appropriate in the spirit of the ITC/ATP Technology-Based Assessment guidelines that were conceptualized to specifically maintain validity, fairness, accessibility, privacy, and security in the digital assessment ecosystems.

### 5.4 GenAI integrity and security: affordances of misconduct and reinvented rules of inference

Theme 4 (Integrity and security) signifies a quickly developing context: generative AI increases the magnitude and delicacy of cheating, coaching, and creating responses, which usual security controls fail in certain situations. This has been the subject of much concern in AI-in-education governance discourses and it has been emphasized that systems should draw a distinction between where support is reasonable and unreasonable and that policy should foresee fresh ways of academic dishonesty. Our results build upon this discussion by demonstrating that stakeholders were not simply demanding stricter surveillance measures; they usually focused on task redesign and inference redesign-reorienting to a type of assessment that can support the assertions of most

people even in an era of high-level AI.

This is consistent with the wider perspective on measurement of security as a component of the validity ecosystem (since security compromised compromises interpretability), which is echoed in the modern-day measurement commentary and technology-based testing instructions. It is also directly related to purpose-first governance: stakeholders suggested that high-stakes uses might demand stronger authenticity evidence (e.g. process evidence, staged evaluations, verification tasks), and that lower-stakes ones might not object to AI as a learning aid. The advice of UNESCO also focuses on human-oriented policy decisions and capacity building, as opposed to a technical containment approach.

### 5.5 Legitimacy, privacy, and institutional readiness: why even working tools do not get adopted

Theme 5 (Legitimacy, privacy, readiness) talks about the essential requirement of adoption: even technically able systems can be used as failure when resided by stakeholders who see them as opaque, invasive, or unfair. This is particularly relevant to Turkey where the privacy compliance and institutional trust do not belong to the category of optional implementation provisions but the conditions of sustainable governance. Law In legal aspects, the Personal Data Protection Law (KVKK, Law No. 6698) explicitly forms the duties and principles of the treatment of personal data and puts a greater value on the fundamental rights, in particular on privacy. The fact that participants focus on transparency, the logic of consent, data minimization and the responsibility of the vendors follow this regulatory trend and the international principles of privacy and security protection in technology-driven tests.

Notably, the stakeholders also conceived legitimacy as a governance result obtained by explainability, appeals process, and visible accountability which are issues that are increasingly put in the regulations. As an example, the EU AI Act establishes common standards on AI and has rules of governance of AI application in high-risk systems (although Turkey is not an EU member, these regulatory developments frequently affect the global best practice and vendor behaviour). The joint effect is that Turkey-specific standards should also contain readiness requirements, such as skills, audit capacity, procurement protections, routines of change-management, in that, absent capacity, then responsible AI will not be operational but performative.

### 5.6 Synthesis: the combination of the themes in response to the research problem

Throughout the themes, the research provides a consistent explanatory framework: responsible AI in educational measurement can be created when (1) score application is defined and risk-based, (2) validity assertions are limited and demonstrated, (3) fairness is monitored continuously with audit-able response strategies, (4) integrity threats are tackled by redesigning inference and tasks, and (5) privacy, legitimacy, and institutional preparedness are not an add-on but fundamental measurement conditions. The synthesis serves as a direct counter to new standards efforts like the Duolingo English Test Responsible AI Standards, and is also compatible with Ho's main argument that measurement-specific AI standards need to be framed based on the context of test use and bias and drift over time monitoring.

### 5.7 Theoretical implications

To start with, the research supporting the fact that responsible AI in educational measurement is best viewed as a continuation of the theory of validity and should not be considered an ethics add-on. The themes indicate that stakeholders measured AI by the answers to questions which are fundamentally, validity questions, what the score means, what claims are justified, as well as what consequences follow, which is consistent with argument-based validity as well as with the overall perspective which holds interpretation and use are central to defensible measurement (American Educational Research Association, 2014; Kane, 2013). In this respect, AI is included in the

inferential chain of score meaning, and validity evidence should clearly mention AI-powered processes, constraints, and failure modes instead of assuming that they are the extraneous tools (Ho, 2024; Bulut et al., 2024).

Second, the results contribute to new research that suggests the need to have a context-of-use as well as a temporal orientation to AI-based assessment governance. The insistence on the existence of the so-called purpose-first risk tiering and accountability among the participants is a clear sign of the argument that risks do not exist in all testing programs; they depend on who uses the scores and on what decisions they make (Ho, 2022; Messick, 1989). Similarly, the aspect of fairness as continuous monitoring supports the existing arguments that fairness should be operationalized as continuous lifecycle governance since drift, games, and evolving populations may alter the impacts of subgroups across time (Ho, 2024; Tabassi, 2023). This provides conceptual clarity to the measurement literature in that monitoring is not a discretionary quality improvement but a fundamental part of the validity argument in testing involving AI as an intermediary.

Third, the research supports the emerging synthesis between the validation and the responsible AI models by demonstrating that the issues of stakeholder legitimacy (privacy, transparency, appealability) may be used as measurements conditions in high-stakes situations. This is in line with recent literature that claims that the responsible AI standards must be converted into testable audit-able and documentable governance practices that can be connected to test equity and quality (Burstein & LaFlair, 2024; Burstein, 2025). This agenda is reinforced by Turkey-based evidence, which focuses on the idea that it is through the governance design that legitimacy is co-produced, but not by technical accuracy.

### 5.8 Implications on practice and policy

To assessment organizations and universities in Turkey, the results have the implication that the adoption strategies must start with their organized use-case taxonomy (high-stakes versus lower-stakes) and apply proportional controls that escalate with decision consequences. This solution aligns with professional testing provisions that provide higher evidence and fairness safeguards in case of greater consequences (American Educational Research Association, 2014; Kane, 2013). In practice, Theme 1 may be converted into governance architecture by the institutions defining roles of accountability (validity lead, fairness/monitoring lead, security lead, privacy lead) and developing documented decision rights regarding changing vendors and updating models and exception handling.

To achieve validity and practice fairly, the findings suggest that institutions ought to take documentation as a common evidence artifact: stating where AI is applied, what assumptions are to be made, what are the known restrictions, and what interpretations are allowed or disallowed. This is in line with technology-based assessment guidelines, which focus on transparency, security, privacy, and adequate use throughout the lifecycle (International Test Commission & Association of Test Publishers, 2022; Salamanca et al., 2025). Theme 3 suggests that fairness should be introduced by practical monitoring procedures - regular subgroup audits, drift metrics, escalation thresholds, and pre-determined corrective measures - that are aligned with the principles of lifecycle risk management in AI governance (Tabassi, 2023, UNESCO, 2023).

To be able to be honest and secure, the results indicate that the institutions cannot afford total reliance on the control mechanisms that can be based on surveillance alone. Instead, they must focus on redesigning assessment, which maintains defensible inferences in the wide availability of AI, including staged verification, process evidence, authentic tasks, and redesigned response settings. This is consistent with up-to-date alerts that generative AI is capable of polluting constructs and compromising the ability to make score comparisons unless the test design and inference rules are revised (Bulut et al., 2024; Ho, 2024). Theme 5 suggests that compliance and trust are to be built in the system design: reducing data, explaining lawful processing reasons,

specifying retention, holding the vendor responsible, and offering an open line of communication and avenues of appeal. The steps are especially relevant considering the Personal Data Protection Law of Turkey (KVKK, Law No. 6698) and the overall trend towards the regulation of AI, where education-related AI is progressively considered to have high impact and require a high level of governance.

### 5.9 Limitations

This research has its limits to interpretation. First, the results are grounded in qualitative narratives and thus they represent the perceptions of the stakeholders with regard to the opportunities of AI, threats of AI, and limitations to feasibility instead of the actual performance of a system or quantified bias. The research is more depth and contextually specific, which increases the explanatory power but restricts statistical generalization; thus transferability is dependent on the readers to determine the similarity of the context to other systems of assessment (Lincoln & Guba, 1985). Second, the sample was created to be as varied as possible in terms of roles and institutional context but still, self-selection is also possible since the people who might be most interested in the issue or most interested in AI, in general, are those who are more likely to participate and salience of some risks may be predetermined.

Third, since the research is on governance and standards development, it does not supplant technical validation research which would be required to test specific models, scoring engines, or proctoring systems in Turkey. Lastly, the analysis is contextual to the high-stakes context and legal framework of Turkey, which enhances local applicability, but might need adjustment prior to the framework application to the contexts with different assessment culture, infrastructures, and regulatory regimes (Ho, 2022; Tabassi, 2023).

### 5.10 Future research directions

The framework generated by this paper should be implemented and tested in the future. One of the most significant follow-ups is to pilot the proposed standards through one or more of the Turkish assessment programs and test whether the governance routines (documentation, version control, subgroup audits, drift monitoring) can be used to enhance defensibility and reduce risk over time, in direct response to the calls to continuously monitor bias and scale drift (Ho, 2024; International Test Commission and Association of Test Publishers, 2022). The corresponding work must establish effective measures and dashboards used in ongoing monitoring corresponding to indicators of fairness, which puts measurement practice into alignment with lifecycle risk governance (Tabassi, 2023; Burstein & LaFlair, 2024).

Second, the stakeholder inclusion should be extended to the wider audience of future studies to explore how the legitimacy, privacy, and fairness are perceived by the students, teachers, and administrators, in particular, in AI-proctored or AI-assisted assessment to demonstrate the actual effect of public trust and acceptance on real-world validity and compliance behaviors (American Educational Research Association, 2014; UNESCO, 2023). Third, the cross-national study with similar high-stakes ecosystems would help to explain which governance factors are universal and which are context-specific, especially concerning the impacts of language, inequities in access, and institutional capacity limitations (Bulut et al., 2024; Ho, 2022).

Lastly, studies need to investigate assessment design creativity in the presence of generative AI, including experimental testing of strategy of task redesign that can maintain construct meaning and be scaled at any cost. This work could also go beyond detection-and-surveillance models to inference redesign which captures the actual working of learners in AI-rich settings (Ho, 2024; UNESCO, 2023).

### 5.11 Conclusion

This paper offers evidence based on the context that is grounded in ways of how Turkish

assessment stakeholders perceive the opportunities and threats of AI in the educational measurement and what they believe it takes to be able to adopt it responsibly. The five themes can be used collectively to suggest a consistent logic of governance: responsible AI in evaluation needs the risk tiering based on purpose first, accountability, the validation based on the focus on validity, and limitations and documentation, fairness through continuous monitoring, strategies of integrity appropriate to generative AI, and legitimacy based on privacy adherence and the readiness of institutions. These results reinforce existing arguments about measurement that AI will increase the predictable uses of artificial intelligence unless it is governed by situational context and maintained by continuous surveillance (Ho, 2024; Bulut et al., 2024). The study provides a viable basis of a Turkey-specific Responsible AI framework that ensures defensible score use without limiting responsible innovation in the assessment practice, by translating the definition of stakeholders into the domains of standards relevance and implementation implications (American Educational Research Association, 2014; International Test Commission and Association of Test Publishers, 2022).

## 6 References

American Educational Research Association. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aruğaslan, E. (2025). Artificial intelligence-based proctored online exams: A study on the experiences of distance education students. *Journal of Buca Faculty of Education, 65*, 2728–2748. https://doi.org/10.53444/deubefd.1543471

Aydemir, M., Erdogdu, E., & Ucar, H. (2024). Exploring online learners' perspectives in relation to proctored exams. *Turkish Online Journal of Distance Education, 25*(4), Article e334. https://doi.org/10.17718/tojde.1459600

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). Routledge. https://doi.org/10.4324/9781315871493-8

Bowen, S. A. (2015). Exploring the role of the dominant coalition in creating an ethical culture for internal stakeholders. *Public Relations Journal*, *9*(1), 1–23.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Braun, V., & Clarke, V. (2024). Supporting best practice in reflexive thematic analysis reporting in *Palliative Medicine*: A review and introduction to RTARG. *Palliative Medicine, 38*(6), 608–616. https://doi.org/10.1177/02692163241234800

Braun, V., & Clarke, V. (2025). Reporting guidelines for qualitative research: A values-based approach. *Qualitative Research in Psychology, 22*, 399–438. https://doi.org/10.1080/14780887.2024.2382244

British Educational Research Association. (2024). *Ethical guidelines for educational research* (5th ed.). British Educational Research Association.

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *arXiv*. https://doi.org/10.59863/MIQL7785

Burstein, J. (2025). *The Duolingo English Test responsible AI standards*. Duolingo.

Burstein, J., & LaFlair, G. T. (2024). Where assessment validation and responsible AI meet. *arXiv*. https://doi.org/10.32038/ltrq.2025.50.09

Burstein, J., LaFlair, G. T., Yancey, K., von Davier, A. A., & Dotan, R. (2024). *Responsible AI for test equity and quality: The Duolingo English Test as a case study*. ArXiv.

Creswell, J. W., & Poth, C. N. (2018). *A Book Review: Qualitative Inquiry & Research Design: Choosing Among Five Approaches*. SAGE Publications. https://doi.org/10.13187/rjs.2017.1.30

Denzin, N. K. (1978). *The research act* (2nd ed.). McGraw-Hill.

European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis. *International Journal of Qualitative Methods, 5*(1), 80–92. https://doi.org/10.1177/160940690600500107

Grant, S., & Khodyakov, D. (2025). Proposal for a critical appraisal tool for Delphi studies. *BMJ, 391*, Article e084509. https://doi.org/10.1136/bmj-2025-084509

Hasson, F., Keeney, S., & McKenna, H. (2025). Revisiting the Delphi technique. *International Journal of Nursing Studies, 168*, Article e105119. https://doi.org/10.1016/j.ijnurstu.2025.105119

Ho, A. D. (2022). Specifying the three Ws in educational measurement. *Journal of Educational Measurement, 59*(4), 418–422. https://doi.org/10.1111/jedm.12355

Ho, A. D. (2024). Artificial intelligence and educational measurement. *Journal of Educational and Behavioral Statistics, 49*(5), 715–722. https://doi.org/10.3102/10769986241248771

International Test Commission, & Association of Test Publishers. (2025). *Guidelines for technology-based assessment* (Version 1.1).

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring. In V. Yaneva & A. von Davier (Eds.), *Advancing natural language processing in educational assessment*. Routledge.

Kallio, H., Pietilä, A. M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, *72*(12), 2954–2965. https://doi.org/10.1111/jan.13031

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kitchen, H., Bethell, G., Fordham, E., Henderson, K., & Li, R. R. (2019). *Student assessment in Turkey*. OECD Publishing. https://doi.org/10.1787/5edc0abe-en

Lane, S. (2023). Validity, fairness, and technology-based assessment. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 127–141). Routledge. https://doi.org/10.4324/9781003278658-11

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.

Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021). Fairness evaluation of automated scoring. In I. Roll et al. (Eds.), *AIED 2021* (pp. 255–267). Springer. https://doi.org/10.1007/978-3-030-78292-4_21

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis. *International Journal of Qualitative Methods, 16*, 1–13. https://doi.org/10.1177/1609406917733847

OECD. (2019). *Student assessment in Türkiye*. OECD Publishing.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & management*, *42*(1), 15–29. https://doi.org/10.1016/j.im.2003.11.002

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). Sage.

Salamanca, S. L. C., Oliveri, M. E., & Zenisky, A. L. (2025). Advancing good practices in a global digital future: ITC/ATP guidelines. *International Journal of Testing, 25*(2), 194–211. https://doi.org/10.1080/15305058.2025.2490234

Salamanca, Y. C. (2025). AI in educational measurement: A practitioner's perspective. *Journal of Educational Measurement, 62*(2), 363–369.

Saldana, J. (2021). *The coding manual for qualitative researchers* (4th ed.). SAGE.

Saunders, C. (2018). *How to teach, lead, and live well: A qualitative in-depth interview study with eight North Carolina teacher-leaders who flourish* [Doctoral disseation Columbia University]. ProQuest and Dissertations.

Shaw, S. (2011). *Tracing the evolution of validity in educational measurement*. Cambridge Assessment.

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*, *22*(2), 63–75. https://doi.org/10.3233/EFI-2004-22201

Squires, A. (2009). Methodological challenges in cross-language qualitative research: A research review. *International Journal of Nursing Studies*, *46*(2), 277–287. https://doi.org/10.1016/j.ijnurstu.2008.08.006

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. https://doi.org/10.6028/nist.ai.100-1

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). Automated scoring framework. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wutich, A., Beresford, M., & Bernard, H. R. (2024). Sample sizes in qualitative research. *International Journal of Qualitative Methods, 23*, 1–15. https://doi.org/10.1177/16094069241296206

Young, J. W., So, Y., & Ockey, G. J. (2013). *Guidelines for test development*. ETS.