*ets research institute

APRIL
2025
RR–25-03

RESEARCH REPORT

Responsible AI for Measurement and Learning: Principles and Practices

AUTHOR

Matthew S. Johnson

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

Responsible AI for Measurement and Learning: Principles and Practices

Matthew S. Johnson

ETS Research Institute, ETS, Princeton, New Jersey, United States

Executive Summary

In the evolving landscape of artificial intelligence (AI) within measurement and learning, there is an urgent need to ensure responsible usage. This document presents ETS's principles for harnessing AI in ways that prioritize ethical considerations, fairness, transparency, and educational integrity. The principles are grounded in the synthesis of widely recognized principles and guidelines from leading organizations such as the National Institute of Standards and Technology, the U.S. Department of Education, OECD, The European Commission, UNESCO, the American Psychological Association, the American Education Research Association), the National Council on Measurement in Education, and the International Test Commission.

At ETS we recognize AI's dual impact—its potential to enhance educational experiences and its inherent risks. We highlight the importance of integrating ethical and sustainable practices throughout all stages of AI implementation, from initial development through post-deployment monitoring and refinement. The principles approach extends beyond promoting fair and equitable educational impacts; it is equally vigilant about addressing AI's environmental implications.

This document not only serves as a guideline for us at ETS; it is also designed to contribute meaningfully to the broader AI, education, and educational testing communities worldwide. By transparently sharing these principles, we aim to foster responsible AI practices that unlock the potential of these technologies, enriching lives from education to the workforce while safeguarding against potential perils. The approach reflects our goal to harness AI for good—helping individuals throughout their lifelong journey of learning.

Acknowledgments

I would like to extend my sincere gratitude to Randy Bennett, Ikkyu Choi, Ayleen Gontz,

Tanner Jackson, Andrew McEachin, Dan McCaffrey, Kara McWilliams, Amit Sevak, and Diego Zapata-Rivera for their insightful comments and thoughtful feedback on earlier drafts of this work.

Abstract

The rapid proliferation of artificial intelligence (AI) in educational measurement presents both transformative opportunities and complex ethical challenges. This paper articulates foundational principles for the responsible integration of AI in measurement and learning, drawing on established guidelines set forth by leading organizations such as NIST, OECD, UNESCO, the U.S. Department of Education, and others. We propose a principled framework encompassing fairness and bias mitigation, privacy and security, transparency, explainability, accountability, educational impact and integrity, and continuous improvement. Through the synthesis of current research, best practices, and cross-sector standards, we highlight practical measures to ensure that AI-driven assessment systems are equitable, valid, and reliable. Special emphasis is placed on the significance of representative data, ongoing bias analysis, secure-by-design development, and stakeholder involvement throughout the AI lifecycle. This approach is designed to foster trust, uphold educational values, and safeguard individual rights. By emphasizing ethical and sustainable practices, we advocate for a vision of AI as a driver of human development—supporting learners, educators, and society at large in the pursuit of educational and economic mobility. The principles and recommendations outlined here offer guidance not only for our organization but serve as a resource for the broader educational and measurement community, charting a course for responsible AI innovation that advances both the science and the practice of measurement in support of lifelong learning.

Keywords: artificial intelligence, AI, fairness, best practices, bias, machine-based learning, assessments, human-in-the-loop, lifelong learning, human development, testing, assessments

Corresponding author: Matthew S. Johnson, Email: msjohnson@ets.org

1 Introduction

An artificial intelligence (AI) system, as defined by the OECD, is a “machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical

or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment" (Russell et al., 2023, OECD Countries section). They also differ widely in their transparency and explainability, features that are increasingly recognized as essential for building trust and understanding in educational contexts. The OECD definition is significant because it underscores the broad applicability of AI across various domains, including education.

As digital tools for learning and assessment evolve, they increasingly incorporate AI technologies. These innovations are designed to offer personalized learning experiences, deliver more insightful feedback, streamline interactions, and increase scalability. Capabilities such as generative AI, adaptive algorithms, interactive tasks, automated scoring, and multimodal data collection can transform the way we capture, process, and interpret educational data. This transformation enables stakeholders—including learners, educators, and policymakers—to access richer, more informative data that can significantly impact educational and career pathways. However, with AI's growing influence comes the responsibility to acknowledge that these algorithmic decisions can profoundly impact human lives. As AI-generated inferences become more central in shaping academic outcomes and future opportunities, it is vital that their deployment is guided by the principles of fairness, transparency, and integrity.

While AI introduces potential to the measurement of human capabilities, significant challenges still need to be overcome to realize this potential responsibly. Central to the goal of responsible use of AI is ensuring that AI aids in fair and equitable assessments, devoid of algorithmic bias, and that it upholds rigorous standards for data privacy, transparency, and accountability. Addressing these challenges begins with thoughtful design from the start, encompassing ethical data collection, robust model building, and comprehensive validation across diverse groups to eliminate unintended biases. Beyond deployment, a commitment to continuous monitoring and adaptation is vital, ensuring that AI-driven measurements remain relevant and equitable in capturing the knowledge, skills, and abilities essential for lifelong learning journeys. It is through such responsible practices that AI can truly contribute to advancing the science of measurement in empowering human progress.

We recognize AI's potential to enhance the measurement of human capabilities by increasing efficiency, scalability, and quality. Our goal is to advance measurement science responsibly, aware of AI's challenges and risks. This document outlines our principles and best practices, based on guidelines from various organizations and our own research findings.

Our hope is that the principles and practices we have developed for AI usage extends its influence beyond the boundaries of our organization. Rather than solely serving as a reference for our stakeholders, partners, and customers, we envision it becoming a valuable resource for the broader AI community, encompassing education, lifelong learning, and human development globally. This approach underscores our dedication to addressing AI challenges responsibly and realizing its potential to enhance the measurement and advancement of knowledge, skills, and abilities throughout an individual's learning journey.

2 Our Principles for the Responsible Use of AI for Measurement

In pursuit of our mission to advance the science of measurement to drive human progress, we are steadfast in our goal to guiding the responsible use of AI within educational contexts and beyond, throughout the lifelong learning journey. Our principles are informed by a thorough examination of established global guidelines, which are detailed in the appendix. The goal of the principles is to improve the way we measure human capabilities through application of AI while upholding individual rights and earning the trust of all stakeholders. By focusing on these principles, we aim to empower individuals as they navigate their learning pathways, positioning AI as a tool to develop skills, ensure fairness, and maintain transparency in all applications.

These core principles are not unique to AI but reflect long-standing values that have guided educational and psychological measurement for decades. At ETS, we have maintained our commitment to these principles through multiple waves of technological innovation—adapting our practices to responsibly explore the potential of each new tool. This history of using ethical, fair, and scientifically validated practices informs our approach as we integrate AI into measurement and learning today.

- **Fairness and Bias Mitigation:** A central principle in applying AI to measurement is ensuring fairness, which is an integral part of advancing human progress. Ensuring fairness in AI requires rigorously identifying and mitigating biases to provide all learners with trustworthy and impartial assessments. The design and deployment of AI systems should actively evaluate and minimize biases, ensuring that assessments are inclusive and accurately reflect the abilities of all individuals, regardless of backgrounds. By committing to these actions, we aim to move closer to a future where educational opportunities are expanded for all, supporting personal growth along every learner's

journey.

- **Privacy and Security:** Safeguarding privacy and ensuring robust security are foundational to using AI responsibly in measurement. Protecting personal data is essential to fostering a trusted learning environment where individuals feel secure in their pursuit of personal development. AI systems must be designed and implemented with strong privacy protections and security measures, ensuring that data is used responsibly and ethically. By prioritizing the privacy and security of all learners, we lay the groundwork for advancing human progress through technology that respects individual rights and supports lifelong learning.
- **Transparency, Explainability, and Accountability:** Transparency, explainability, and accountability are pivotal in building trust and confidence in AI systems used for measurement. It is essential that stakeholders understand how AI-driven decisions are made and have access to clear explanations of the outcomes. By ensuring that AI operations are transparent and accountable, we empower learners, educators, and all stakeholders to engage confidently with these technologies. This openness and responsibility not only facilitates informed decision-making but also fortifies the foundation for advancing human progress by promoting trust and integrity in AI throughout the lifelong learning journey.
- **Educational Impact and Integrity:** AI should be employed in a way that supports educational objectives and upholds integrity. For AI to be a true catalyst for advancing human progress, it needs to make valid and accurate inferences about learners' skills and knowledge. The design and application of AI in measurement should align with personal development goals, respecting individual rights and privacy, while considering potential unintended effects. By doing so, we ensure that AI positively impacts educational experiences and supports individuals' lifelong learning journeys.
- **Continuous Improvement:** The use of AI in measurement should be subject to ongoing evaluation and refinement to ensure they remain effective, relevant, and responsive to the evolving needs of learners. By continually assessing and improving these systems, we adapt to changes in educational contexts and incorporate the latest innovations and research. This commitment to continuous improvement not only improves the accuracy

and utility of AI applications but also contributes to the advancement of human progress by supporting sustained personal growth and learning throughout an individual's life.

Although these principles apply broadly to all AI technologies used in measurement and learning, they are particularly pertinent to the rapidly evolving class of generative AI systems. The unique capabilities and risks associated with generative AI—including the generation of new content, interactions with learners, and potential for unanticipated outputs—underscore the importance of a robust, principled approach to responsible AI.

In the sections that follow, we explore these core principles further, exploring their application for the measurement of human capabilities. In doing so, we hope to provide a foundational guide for anyone seeking to integrate and leverage AI technology in measurement, with a strong commitment to fairness, privacy, transparency, and positive educational impact. However, they do not represent an exhaustive list; ongoing research and technological advances will continue to shape and expand upon these recommendations.

3 The Principles and Practices for Fairness and Bias Mitigation

Among the guiding principles for the use of AI in education, fairness and bias mitigation are of particular importance (Baker & Hawn, 2022). The principles of fairness and bias mitigation work toward avoiding any discriminatory actions or decisions, promoting social and economic mobility, and ensuring integrity in the learning and assessment process.

The use of AI for the measurement of human capabilities should promote fairness by eliminating any factors that could advantage or disadvantage learners based on their socioeconomic status, cultural background, race, gender, disability, or any attribute other than the construct of interest. For example, an AI scoring algorithm should not favor one group of students over another due to their socioeconomic status, culture, race, or gender beyond what might be explained by differences in the construct.

Developers should take proactive approaches to mitigate the potential for bias due to AI-supported measurement, learning, and development applications. Biases can originate from various sources, including training data, algorithms, or decision-making processes. It is essential that the developers of AI tools understand potential biases and take steps to mitigate them to achieve fairer outcomes.

To mitigate biases and causes of unfairness, procedures should be established including:

- **Representative Data:** Ensure data used to train AI systems represents the full spectrum of learners' backgrounds, abilities, and experiences. This will help to improve the fairness and generalizability of the AI outputs.
- **Bias Analysis:** Regularly conduct analyses to detect and rectify any biases in the AI algorithms' outputs, maintaining fairness across demographic groups.

Promoting fairness and mitigating bias are integral to the responsible use of AI for the measurement and development of human capabilities. In the sections below we share some best practices for each of these areas.

3.1 Best Practices for Data Representativeness

Developers of AI applications for the measurement of human capabilities should ensure that data used to train the algorithms is representative of the population on which it will be applied (Clemmensen & Kjærsgaard, 2023). Data representativeness has many interpretations. In fact, Kruskal and Mosteller devoted four articles to exploring the different notions of the concept (Kruskal & Mosteller, 1979a, 1979b, 1979c, 1980). Therefore, it is crucial to consider how the data used to train and evaluate AI-driven measurement systems can impact their performance. Some best practices associated with ensuring data used to train and evaluate AI systems are representative include the following:

- **Define the Population:** Clearly define the population that the AI system will serve. This definition should include essential demographic characteristics such as age, educational level, socioeconomic status, cultural and linguistic background, geographic location, and any other factors that are relevant to the educational context in which the AI will be used. For example, if an AI system is being developed for English language learners, then the population might include individuals from many different native languages and socioeconomic backgrounds.
- **Define the Variables:** Fairness in machine learning and AI is often defined in terms of interrelationships among three sets of variables: (a) the output/prediction of the AI system; (b) the human decision/target that the AI is trying to replicate; and (c) the indicators of the groups for which we want to ensure fairness; these might include demographic groups like sex, urbanicity (e.g., urban/rural), or socioeconomic status

(Barocas et al., 2023). Defining these variables transparently is important to minimize any risk of misinterpretation of the fairness evaluation.

- **Diversify Data Collection:** Ensure that the data collected and used to train and evaluate the fairness and bias of the AI system adequately represents the population on which the AI system will be applied. In particular, ensure that student groups of interest (e.g., group defined by gender, ethnicity, socioeconomic status, educational background, etc.) are adequately sampled so that fairness and bias can be evaluated on these groups.
- **Capture Contextual Variables:** Data collection should also capture relevant contextual variables relevant to the learning environment that might be related to the performance of the AI system, such as their linguistic background, disability status, or access to educational or training resources.
- **Be Mindful of the Demographic Makeup of the Data:** Understanding the demographic composition of the data is crucial to avoid introducing bias. In the case of an AI system designed for educational settings, this could mean considering the representation of students of different backgrounds. For instance, if the data overwhelmingly includes test scores from suburban middle-class students, the AI system might not perform as effectively when applied to students in underprivileged urban schools or in schools in rural areas. Similarly, if the data does not include enough observations from students with special education needs, it might provide less accurate predictions for these students. Any imbalance might impact the training and evaluation of the AI system and influence the interpretation of fairness and bias analyses. Therefore, it is important to weight samples as needed to ensure data used for training and testing AI models is representative of the population it is intended to serve.
- **Continuous Data Updates:** Regularly update the data used to train and evaluate AI systems to reflect changes in the student population, adjustments to educational standards, or shifts in societal factors. For instance, if using an AI chatbot in a conversational-based assessment, updates might include incorporating new language usages, slang, or idiomatic expressions common among current student cohorts. Additionally, ensure the AI understands and assesses evolving curriculum topics or reflects recent advancements in technology and pedagogical strategies. This ongoing

refinement will help ensure that AI models continue to provide fair and relevant results.

Though adopting best practices for data representativeness significantly minimizes bias in AI systems utilized in education and assessment, it is important to understand that this is just one part of a multifaceted issue. A continuous commitment to fairness and bias reduction is required at each stage of the AI lifecycle. This commitment extends from how we collect data and devise algorithms to how we evaluate performance and make decisions based on the AI's output. To put it simply, having representative data is merely a starting point; the important task of bias mitigation and fairness assurance spans the entire development and deployment of AI systems.

3.2 Best Practices for Fairness and Bias Analysis

Bias can undermine the fairness and validity of AI systems in educational measurement and learning applications. To address this, we outline the following best practices for bias analysis.

- **Explicitly Define Fairness and Bias in Context:** Stakeholders should document their understanding of fairness and bias in the particular context of the educational setting in which the AI system will be applied. Fairness might imply different treatment of different groups to adapt to their specific needs. In other settings, differential treatment could be seen as unfair. In fact, the Joint Standards (AERA et al., 2014) discusses four views of the concept of fairness: (a) fairness in treatment during the testing process; (b) fairness as lack of measurement bias; (c) fairness in access to the construct as measured; and (d) fairness as validity of individual test scores.

  Similarly, even when only considering the view of fairness as lack of bias, the machine learning and statistics literature have multiple definitions of fairness (Barocas et al., 2023; Carey & Wu, 2023; Castelnovo et al., 2022), so considering which one is most appropriate for a given application is important. Barocas et al. (2023) consolidates the many definitions of lack of bias into three forms of fairness. The three forms of fairness are defined in terms of the sensitive attribute being considered (e.g., race, gender), the target variable (e.g., true score), and the prediction/output of the model:

  - *Independence:* Independence fairness requires that the prediction/output be statistically independent of the sensitive attribute.

- *Separation:* When the sensitive attribute and the target are associated with one another, independence may not be appropriate. Separation fairness is when the sensitive attribute is conditionally independent from the AI output given the target variable. Separation requires that any variation we see in the AI output across groups be associated with true differences in the target.
- *Sufficiency:* Sufficiency fairness uses a different conditional independence property, namely that the sensitive attribute is conditionally independent of the target given the output of the AI. This type of fairness suggests that if we have the output of the algorithm, knowledge of the sensitive attribute provides no further information about the target we are trying to predict.

In the context of educational testing, differential item functioning (Holland & Thayer, 1998; Holland & Wainer, 1993), differential algorithmic functioning (Suk & Han, 2023), and the fairness metrics of Johnson et al. (2022) and Johnson & McCaffrey (2023) are all aligned with the definition of separation fairness.

Other practices to consider when defining fairness and bias include the following:

- **Involve Stakeholders:** Engage stakeholders to define fairness and address bias in educational measurement and learning contexts.
  - *Learners:* Learners can offer key insights into perceived fairness or bias in AI systems and how these impact their learning experiences and outcomes.
  - *Educators:* They can shed light on fairness in terms of educational inputs (e.g., variations in learning materials or teaching techniques), as well as any perceived differential effects on student engagement or performance due to the AI system.
  - *Administrators:* They can provide an institutional view on fairness, relating to school policies and goals, resource allocation, and broader socio-educational issues.
  - *Policymakers:* They can help ensure the broader societal and legal definitions of fairness are embedded in the AI systems and can also address larger educational challenges.
  - *AI Developers and Researchers:* These individuals can contribute by translating

the diverse notions of fairness into practical, quantifiable measures that can be integrated into AI system design and evaluation.

Including these diverse viewpoints helps ensure fairness is defined appropriately, considering the full range of potential biases and impacts on various groups within the educational system.

- **Consider Legal and Ethical Standards:** Definitions of fairness and bias should be aligned with existing laws such as antidiscrimination laws (e.g., the U.S. Civil Rights Act, Individuals with Disabilities Education Act), ethical guidelines established by relevant organizations (e.g., the Joint Standards, UNESCO, etc.), and institutional policies.
- **Specify Acceptable and Unacceptable Biases:** It is essential to clarify what constitutes acceptable and unacceptable bias. Below are examples of each:
    - *Acceptable Biases:* These could be forms of positive bias designed to address existing imbalances. For example, having an AI system adjust its difficulty level to the learner's current ability level can be seen as an acceptable bias. It is intended to help the students learn at their own pace.
    - *Unacceptable Biases:* These biases result in unjust or discriminatory practices. For example, an AI grading system that consistently awards lower grades to essays written by non-native English speakers, without factoring in their fluency level, could be deemed unfair. Any AI system that discriminates based on protected classes (e.g., race, gender, disability) would also be considered biased.

Clear identification of such biases can guide the development of AI systems to prevent harmful discrimination and promote fair practices. The acceptability of certain biases may evolve with societal norms, so continuous scrutiny and dialogue is important.

- **Adopt a Multidimensional Perspective:** Fairness and bias can be multifaceted (Castelnovo et al., 2022) depending on the point of view. It can be useful to adopt various fairness definitions.
    - *Group Fairness:* AI systems must avoid disadvantaging learners based on race, socioeconomic status, or ability. For instance, an AI tool for college admissions

should not consistently recommend fewer offers to qualified students from marginalized communities.

- *Individual Fairness:* Similar learners should receive similar treatment. For instance, two individual learners who have made similar progress in a personalized learning environment should receive similar recommendations for the next steps or learning resources, without favoring one learner over the other.
- *Counterfactual Fairness:* An AI system's decisions remain the same if we hypothetically change a sensitive attribute, while all other factors remain equal (Kusner et al., 2017). For example, if we change the socioeconomic status of a student (leaving all else the same) in an AI system designed to recommend students for a competitive internship program, it should not result in a different recommendation.
- By considering these various dimensions of fairness, educational AI tools can be designed and evaluated for comprehensive fairness and avoid perpetuating existing inequalities.

Definitions of fairness and bias may evolve with societal norms and values, technological advancements, and increased understanding of AI potential and pitfalls. Continuous reflection and dialogue about these concepts should be encouraged.

- **Integrate Fairness and Bias Analysis in Design:** The integration of bias analysis into the developmental stages of AI systems can prevent potential harm to learners and promote fairer outcomes. It requires meaningful engagement with diverse stakeholders throughout the AI lifecycle. This integration starts by mapping all affected communities and ensuring their representation in design. Co-defining fairness with these communities, collecting representative data, and empowering authentic, ongoing participation—rather than token consultation—are all essential.

  For example, when designing AI systems for conversation-based assessments, developers should work with diverse stakeholder groups to anticipate biases that may favor specific dialects, accents, or communications styles, potentially disadvantaging speakers who use nonstandard forms or have different cultural communication norms. Similarly, AI-scoring algorithm developers should work with stakeholders to anticipate

potential biases that might favor certain writing styles or penalize based on common writing traits of English language learners. A proactive approach to fairness helps to uncover hidden biases before they can cause harm to learners.

- **Use Comprehensive Evaluation Methods to Assess Fairness and Bias:** Incorporate a wide array of evaluation methods, both quantitative and qualitative, to assess fairness and detect bias in AI systems. This should include psychometric, statistical, and machine learning techniques. It might include other methods, such as examining the congruence between the target construct and its representation encoded in the AI system. The method used should align with the specific definitions of fairness and bias being mitigated but may include techniques such as differential item functioning for test content, disparate impact analysis or demographic parity for system outcomes (Aigner et al., 2024; Miao & Gastwirth, 2013), and fairness-aware algorithms (Pan et al., 2021) in the modeling process.
- **Employ Bias Remediation Strategies:** If bias is detected, implement steps to remediate it. For example, an AI model might be trained to predict future student or employee performance. A biased model may unfairly predict lower performance for learners from certain demographic groups, such as students from rural locales. Remediating this issue might require adjusting the algorithm using strategies such as these:
    - *Cost-Sensitive Learning:* This can be implemented to weigh the misclassification of certain groups more heavily (Elkan, 2001). For instance, in an educational setting, a higher cost could be assigned to false negatives for learners from rural communities. This would make the AI model work harder to correctly classify these students, thus reducing biases against these students.
    - *Constrained Learning:* Constrained learning can be used to ensure that the model's predictions meet certain fairness constraints. For example, Johnson & McCaffrey (2023) demonstrated how penalization methods could be used to ensure that separation unfairness metrics are zero in the training sample in an automated scoring algorithm.
    - *Penalized Learning Methods:* Penalized methods penalize algorithms for producing unfair results during the training stage. Depending on how strong the

penalty is, the resulting algorithm can range from the original algorithm (no penalty) to a fully constrained algorithm (infinite penalty). Thus, the amount of penalization can be tuned to improve the fairness of the algorithm while maintaining acceptable levels of accuracy. Yao et al. (2019) and Johnson & McCaffrey (2023) demonstrate this approach in the context of automated scoring.

It is important to note that ethical considerations come into play when using these techniques. Improving the group-level fairness for one set of demographic groups might reduce the fairness for another set of demographic groups. Similarly, improving group-level fairness might reduce individual-level fairness (Castelnovo et al., 2022). Therefore, any remediation strategies must be applied with care and awareness of potential unintended consequences. The use of AI for measurement of human progress requires a balanced approach, providing opportunities for all learners while recognizing the complexities of fairness across diverse contexts and individual needs.

- **Evaluate the Fairness-Accuracy Trade-Off:** Removing bias might reduce the performance of the AI system in terms of accuracy. Evaluate the "fairness-accuracy trade-off" (Corbett-Davies et al., 2017) and aim for a balance: a system that is as fair and as accurate as possible (Buijsman, 2023). For example, when creating an AI tool designed to assess skills within career development programs, adjustments to mitigate bias might lead to a reduction in predictive accuracy. The objective is to maintain a system that effectively evaluates skills while ensuring all individuals are measured fairly, supporting personal growth and economic mobility for all learners.
- **Report and Document Bias Analysis Results:** Systematically report and document the process and results of each bias analysis. This will help track progress over time and identify recurring or persistent issues. For example, in a skills assessment tool used for career development, regular documentation of bias analyses can uncover patterns in assessments that disadvantage certain groups, guiding necessary adjustments and increasing stakeholder confidence in the tool's fairness.

It is important to remember that bias analyses are an ongoing process, not a one-time task. Regular and systemic bias analyses are vital to ensure fair AI applications in education.

4 The Principles and Practices of Privacy and Security

Incorporating AI into measurement settings necessitates a commitment to privacy and security. These principles are vital for safeguarding personal information and guaranteeing the integrity of AI systems. By promoting these principles, we create an environment of trust that enables AI technologies to operate effectively and ethically.

The core principles of privacy and security that guide our implementation include the following:

- **Secure by Design:** Ensure AI systems are designed with built-in security measures to safeguard against potential threats and vulnerabilities from the outset.
- **Secure Development and Deployment:** Adopt and maintain rigorous security practices throughout the development and deployment process to protect AI systems against unauthorized and malicious use.
- **Data Protection and Privacy:** All personal data collected, processed, and stored by AI systems should be handled in a manner that respects privacy rights and complies with relevant laws and regulations. Basic practices for data protection and privacy include obtaining informed consent, anonymizing identifiable information, and ensuring strict access controls to sensitive data.

In addition to these principles, it is also important to continuously monitor and review the effectiveness of security measures and to update them as necessary in the face of evolving threats and risks. This includes regular auditing and testing of systems to identify and fix any security vulnerabilities.

Implementing strong privacy and security measures not only protects organizations and their stakeholders but also helps to build trust and promote the responsible use of AI in measurement contexts.

4.1 Best Practices for Secure by Design

Secure by design (SBD) is a method of designing technologies that are naturally resistant to cyber threats. The U.S. Cybersecurity Infrastructure Security Agency (CISA; 2023c) defines SBD as systems “built in a way that reasonably protects against malicious cyber actors successfully gaining access to devices, data, and connected infrastructure” (p. 8). Central to SBD

is the understanding that the user's or customer's security requirements should be integrated into the technology from its inception (CISA, 2023b).

In the context of educational technologies, particularly ones using AI, SBD is essential. Due to the nature of the data these technologies process—personal details of students and staff, academic records, and potentially sensitive assessment data—they are a high-risk target for cyber threats. Therefore, AI-powered learning technologies should include robust and effective security measures right from their design stage. To further this commitment, CISA (2023a) encourages K12 educational technology suppliers to take a Secure by Design Pledge, which promotes taking responsibility for customer security, demonstrating transparency and accountability, and holding top leadership accountable for cyber security.

Practices that can help integrate SBD principles into the development of learning and measurement applications that rely on AI include:

- **Understand and Anticipate Security Needs:** Developers must proactively understand and anticipate how AI systems will engage with various stakeholders, whether in educational settings—like interacting with students, teachers, and parents—or in skills assessment programs, involving candidates and employers in the hiring process. These systems, used for personalized learning, automated scoring, or skills assessment often capture a wide array of data, including behavioral patterns, personal academic progress, and other sensitive information. Recognizing the scope of these interactions is crucial for determining the security measures needed to protect such sensitive data.

  Consider an AI system used for personalized assessment where each student's learning style and academic strengths and weaknesses are accounted for while designing their unique assessment. In such a system, the AI will continuously collect and analyze students' data, including their personal characteristics, learning patterns, and feedback. The developers must use their knowledge of the system to anticipate potential risks such as unauthorized data access, or misuse of personal data, and thereby devise protective measures accordingly, including robust encryption methods, secure data transmission protocols and strong authentication mechanisms.

- **Model the Threats to Security:** Understanding potential threats is crucial to designing a secure system. Conducting a threat modeling exercise helps identify potential security

vulnerabilities, understand the potential impact of these vulnerabilities, and develop strategies to mitigate identified threats. Developers should consider a variety of threats, including both internal and external threats, and those related to human error. For instance, in a virtual learning environment where students use an AI-based tool to submit homework and take tests, potential threats could span from unauthorized access to a deliberate manipulation of academic results to errors, such as accidentally altering a learner's data.

Understanding these potential risks allows developers to preemptively design security mechanisms, including strong access control measures, secure channels for data transmission, effective encryption, anti-malware tools, systematic data backup, and more.

- **Incorporate Security Controls From the Start:** Integration of strong security measures from the beginning is key to ensuring reliable and trustworthy AI systems for measurement and to support human capability development. Key security controls should include techniques for managing access to the system, protecting data both at rest and in transit, managing authentication and authorization, as well as other measures like error handling, logging and monitoring. The scope and strength of these controls should be proportionate to the sensitivity of the data involved and the potential risks of a security breach.
- **Plan for Ongoing Security Updates:** Security is not a one-time task but an ongoing responsibility. As technology evolves and security threats change, security measures need to be continuously updated. Regular security assessments and updates are necessary to ensure that AI-powered measurement and learning platforms remain secure throughout their life cycle.
- **Educate and Train Staff:** Ensure that all personnel involved in the design, implementation, and maintenance of AI systems are knowledgeable about security considerations and trained in best practices. This helps create a security-focused culture within the organization.
- **Consider Security Benefits and Trade-Offs:** While designing secure AI systems, balance the need for robust security with considerations of usability, cost, and performance. Security measures may increase costs or impact user experience but are

essential for protecting sensitive data and maintaining trust.

Thus far, we have primarily discussed security in terms of protecting against external threats to the AI system, such as unauthorized access to student data. In contrast, security features like those in remote proctoring systems focus on internal risks, such as verifying test-taker identity and preventing cheating. For example, consider an AI-assisted remote proctoring used for administering tests outside of a traditional test setting. For security purposes, the system might incorporate voice recognition technology for student identification and monitor unusual behavioral patterns for indications of cheating. Such security features ensure the right student is taking the test, and it keeps a check on unfair practices. However, these features could also have trade-offs. For example, facial recognition technology may come with additional potential privacy concerns. Similarly, while monitoring tools are important, overly invasive methods may negatively impact the student's testing experience, causing stress or privacy concerns.

These benefits and trade-offs should be carefully considered and balanced. At all times, prioritize security controls that most effectively reduce identified risks and align with the system's overall objective.

By incorporating security into the design of AI systems, developers can ensure that the systems are better protected against potential threats, safeguarding both the integrity of the AI system and the privacy of the data it handles.

4.2 Best Practices for Secure Development and Deployment

The principles of secure development and deployment are focused on integrating security considerations into every stage of the AI system lifecycle, from initial development through to deployment and maintenance. Here are some of best practices for incorporating these principles.

- **Follow Secure Coding Practices:** Encourage developers to write secure code by following the best practices for secure coding. This could include practices like input validation, error handling, logging, and secure use of third-party libraries and application programming interfaces (APIs). For example, rigorous input validation should be used to prevent injection attacks where malicious scripts are inserted into input fields to manipulate test results or compromise sensitive student data. Secure session management can protect against session hijacking attacks that could give unauthorized individuals

access to test questions, answers, or results.

- **Secure the Supply Chain:** The supply chain in the context of AI learning and measurement systems encompasses multiple elements from hardware and software suppliers, third-party libraries and APIs used in development, cloud service providers, data providers, and others. Ensuring security across the supply chain is crucial to prevent vulnerabilities that might compromise the privacy and integrity of educational data. For example, third-party libraries used in developing learning and measurement applications should be kept updated to the latest secure version to prevent security breaches. Similarly, cloud service providers used in storing test-taker data should follow stringent data privacy regulations and best practices to prevent data leaks and unauthorized access.
- **Document AI Assets:** Maintain an inventory of AI models, features, and components used in the system. This helps in managing updates, patches, or bug fixing, and thus contributes significantly to maintaining overall security. For example, an AI system might have multiple AI models performing various functions like automated grading of answers, predicting student performance for adaptation, or flagging potentially problematic behavior. Each of these must be tracked for any necessary updates, patches or potential bugs that may affect overall system security. AI components can also include things like the datasets used for AI model training and evaluation, which should be safeguarded to protect test-taker privacy. Any modifications to AI assets should also be documented as part of ensuring secure deployment.
- **Manage Technical Debt:** Technical debt refers to the future costs incurred because of the decisions made during the development phase, especially those that prioritize quick deployment over best practice. Avoid accumulating technical debt as it could lead to security vulnerabilities. For example, using an understudied or suboptimal algorithm for skills assessment might initially expedite deployment, but could later result in unreliable results and security gaps. When technical debt arises, manage it effectively by keeping track of it, documenting it, and prioritizing its repayment.
- **Automated Security Testing:** Incorporate automated security testing tools that can detect security vulnerabilities as code is being written. This practice will not only help your development team detect and fix issues earlier in the development process but will

also foster a culture of security awareness. For instance, use automated bots to test user access controls in skills assessment platforms, ensuring that only authorized individuals—such as learners and administrators—access sensitive information.

- **Conduct Regular Security Audits and Penetration Testing:** Regular audits and penetration tests are crucial for securely developing and deploying AI systems. They help identify potential vulnerabilities in the system and verify the effectiveness of security controls. For measurement and learning organizations that maintain sensitive learner data, security audits can identify security gaps in how data is stored, transmitted, and accessed. The results of such audits can help organizations prioritize resource allocation to address needed security improvements.
- **Develop Incident Management Procedures:** Develop clear procedures for identifying, responding to, and recovering from security incidents. This involves the following:
    - *Incident Detection:* Implement mechanisms to detect and report anomalies or suspicious activities in the system that could indicate a security breach, such as unusual network traffic or repeated student login attempts.
    - *Incident Response:* Have a response plan in place in case of a security incident. This can include steps to assess the extent of the breach, contain and eliminate the threat, and restore the affected systems to normal operation. It is also essential to thoroughly document each action taken during the response process for future reference and regulatory compliance. In the end, lessons should be learned to enhance the security measures in place and prevent a similar incident in the future.
    - *Incident Recovery:* Develop a plan to recover from a security incident. This will include data backup and recovery procedures, system restoration and patching, and communication plans to inform affected parties.
    - *Post-Incident Analysis:* Once the incident is over and systems are restored, analyze the incident to understand its root cause and impact, and establish measures to prevent similar occurrences in the future.

Consider, for example, an AI powered personalized measurement system.

An effective incident management process might start by incorporating advanced monitoring tools to watch for unusual activity, such as unexpected system behavior, unusual data patterns, or irregularities in results. If a potential breach is detected the system might be immediately taken offline to prevent further harm. If test-taker data was manipulated or exposed, data backup and recovery procedures already in place should be executed, and the system should be corrected to prevent future breaches.

It is impossible to eliminate the risk of security incidents, but robust incident management procedures can minimize their impact and ensure prompt and effective response.

Security is not a one-time task, but an ongoing effort. With a proactive, integrated, and holistic approach to secure development and deployment, the level of protection provided by AI systems in educational testing can be significantly enhanced.

4.3 Best Practices for Data Protection and Privacy

With technological advancements, the amount of data we can collect, store and process has increased significantly. As data play a vital role in AI development and deployment, it has become increasingly important to ensure its protection and respect users' privacy rights. Here are some best practices for data protection and privacy:

- **Safeguard Personal Data:** Protect personal data from unauthorized access, misuse, or loss. Store data using encryption, robust authentication protocols, secure cloud services, or other advanced security mechanisms. Examples of personal data relevant in the context of measuring human capabilities include assessment results, learning progress, demographic information, and user credentials.
- **Collect Only Necessary Data:** Limit data collection to what is strictly necessary for the purpose of the AI application. The less data you collect, the less risk there is. For example, a personalized assessment platform might need to collect information such as the student's name, a class indicator, information about their interests and background for personalization, cognitive response data, and their behavioral interaction (process) data (e.g., time on each question, skipped items, etc.). The system should not collect other information such as their home address, or parents' income unless they are absolutely

necessary.

- **Anonymize or Pseudonymize Data:** If identifying data is not necessary for the functioning of the AI system, anonymize or pseudonymize it. Learners, educators, and other stakeholders may need to see real names and grades in a user interface, but in the backend database, and in the data that is used to train AI algorithms, direct identifiers should be replaced with unique IDs. For example, data that might have originally been stored with "[Student's Name - Grade - School - Teacher - Score]" would be changed to "[Unique ID - Score]." A second, more secure, data location would store information to translate the unique identifier back to the student information. This practice of anonymizing data reduces privacy risks if the data were to be leaked or misused.
- **Consider Federated Learning Approaches:** Recent advances in privacy-preserving machine learning, such as federated learning, allow models to be trained collaboratively across decentralized devices or servers without exposing raw data. When combined with differential privacy techniques, these approaches can further safeguard sensitive information while enabling robust, distributed model development (Wei et al., 2020).
- **Provide Transparency and Obtain Informed Consent:** It is important to provide clear, easily understood explanations of what data is being collected (e.g., performance measures, response time, etc.), how the data is being used (e.g., to tailor instruction, identify learning gaps, etc.), and who has access to the data (e.g., teachers, system administrators, etc.). This transparency helps to build trust and ensure alignment with data protection standards.
- **Implement Strong Access Controls:** Limit access to sensitive data to only the necessary personnel and implement strong authentication protocols to prevent unauthorized access. Students' confidential data, such as test answers, grades and personal information should be restricted to only those who need the information for valid purposes. Robust security measures, such as multifactor authentication and stringent user-access controls can help safeguard against unauthorized access to confidential student information.
- **Comply With Legal and Regulatory Requirements:** Ensure that all data collection, storage, and processing activities comply with all applicable data protection laws and

regulations, such as the Children's Online Privacy Protection Act (COPPA) or the Family Educational Rights and Privacy Act (FERPA) in the United States or the General Data Protection Regulation (GDPR) in the European Union. For example, FERPA requires that student educational data be shared only with individuals who have a legitimate educational interest and proper authorization, and that students (and parents of minor students) be given access to their records, the ability to request amendments, and control over the disclosure of personally identifiable information.

- **Regularly Review and Update Data Protection Measures:** Just as technology and threats evolve, so should your data protection strategies. Regularly review and update them to ensure ongoing effectiveness.

Maintaining privacy and security of data in AI-enhanced educational tools and platforms requires a holistic strategy, which includes deliberate measures undertaken before, during, and after the collection and processing of learner data. Concentrating on data privacy and security not only safeguards learners and their information, but it also fosters trust in the AI-based education applications.

5 The Principles and Practices of Transparency, Explainability, and Accountability

The application of AI in educational and workplace assessment settings can significantly shape individuals' learning outcomes, opportunities, and future prospects. To provide stakeholders with the information they need to make informed decisions and protect their rights, it is essential that AI applications in these settings adhere to the following principles:

- **Transparency:** The development and operation of AI systems in education should be transparent. This transparency should include a clear and open disclosure of how these systems are designed, trained, implemented, and evaluated. Stakeholders, including policymakers, education administrators, teachers, students, parents, employees, and employers should understand how the AI system's operation impacts curriculum design, teaching and training decisions, personalized learning and development experiences, and education and workplace decisions in general. Transparency fosters trust among all stakeholders.
- **Explainability:** The predictions, decisions, or recommendations made by an AI-based

measurement and learning systems should be explainable. Users should be able to understand and make sense of the AI's outputs in a meaningful way. For instance, if an AI system recommends a personalized learning path for a student, it should be able to explain why this path is suitable based on the student's learning style, performance, and preferences. Similarly, in skills assessment, the system should clarify how specific competencies were evaluated and why certain skill development paths are suggested. This is crucial in both educational and professional contexts, where explanations can guide teachers and managers in supporting learning and development, enabling individuals to make informed decisions about their learning journey or career progression.

- **Accountability:** The developers of educational AI applications should be accountable for the systems' decisions and impacts. This involves being able to detect and correct errors, mitigate biases, and rectify harmful impacts if they occur. For example, if an AI system incorrectly assesses a student's performance, there should be mechanisms to correct this and ensure the student is not unfairly disadvantaged. There should be clear lines of responsibility for different aspects of the AI system's operation, and mechanisms for oversight and redress.

Each of these principles is essential for safeguarding the rights and interests of all participants where AI is employed, including employees, students, teachers, policymakers, and parents. They ensure that all users can comprehend and interact with AI applications, granting the opportunity to question and investigate their functionality and decisions. This degree of openness is especially critical for the measurement of human capabilities, where the outcomes can significantly impact an individual's education, career, and social and economic mobility. Ensuring transparency, explainability, and accountability not only fosters trust in these systems but also facilitates their ethical and responsible usage for the measurement of human capabilities.

5.1 Best Practices for Transparency

Transparency refers to the practice of being open, honest, and straightforward about the details of an AI system. This includes aspects such as the data used, the algorithms implemented, the decisions made, and the potential impacts of applying AI systems in education. The goal of transparency is to provide stakeholders with clear, comprehensive, and accessible information so

they can understand and make informed decisions about the use of AI systems. Best practices to ensure transparency in AI measurement applications include the following:

- **Disclose Technical Details:** Details such as the AI model used, dataset properties, performance metrics, and the assumptions made in developing the system should be disclosed to all stakeholders. These technical specifications can guide a deeper understanding of the system functionalities and limitations. In an AI-based plagiarism detection system, transparency can be upheld by disclosing the AI methods used, information about the training and evaluation datasets, as well as performance metrics like accuracy and sensitivity. Key assumptions, for instance, assuming certain behavior patterns as indicative of cheating, should also be communicated to ensure those using the system are aware of potential limitations.
- **Explain Potential Impacts:** Description of potential impacts on individual learners or groups, both positive and negative, helps stakeholders appreciate the possible consequences of using an AI system. This could involve explaining how the AI system might influence learning outcomes, behaviors, or experiences. For example, explaining the potential consequences of using a personalized assessment might include pointing out the positives, such as the ability of such systems to tailor assessments based on each student's unique learning style and understanding. On the other hand, potential drawbacks should also be highlighted, such as the possibility of the system creating an over-reliance on technology or potentially limiting the development of certain skills like critical thinking or problem-solving if not implemented carefully. Also, there could be concerns about data privacy and equity if the technology is not accessible to all students. All these impacts must be thoroughly examined and explained to stakeholders to ensure informed decision-making about the use of AI systems in education.
- **Present Clear Usage Guidelines:** Documenting a clear set of usage guidelines supports stakeholders in understanding how to interact with the AI system effectively. These guidelines should detail how and when to use the system and provide guidance on interpreting and acting on the system's outputs. For example, if an AI system is developed to give formative feedback on an individual's writing ability, it should provide information on how the system works and clear guidelines on how to employ it. Users

should be informed about when and where it is most effective to use the system; maybe the feedback algorithms work best on structured essays or narratives, but less so on poetry or other forms of creative writing. Similarly, in workplace skills assessment, guidelines might specify that the system excels in evaluating technical competencies but requires human oversight for assessing leadership skills.

- **Describe Governance and Oversight Mechanisms:** Transparency also extends to the mechanisms and processes used for governing the use of AI in education. These may include the standards adhered to, the procedures for updating and refining the system, methods for handling complaints or feedback, and safeguards in place to protect users' rights and data privacy. Describing this information to learners, educators and other stakeholders provides a clear understanding of how the AI system is managed and their rights in relation to the system.
- **Use User-Friendly Language and Formats:** Information should be presented in a language and format accessible to diverse stakeholders, such as teachers, students, administrators, and policy makers. Avoid using overly complex terminologies without explanation, and consider the use of visual aids, infographics, videos, or interactive demonstrations to illustrate points. Providing information in user-friendly ways helps ensure that all stakeholders, regardless of their technical background, can understand the workings and implications of the AI system, promoting more meaningful engagement.

Maintaining transparency in AI-powered measurement applications can help to build trust among stakeholders, promote responsible use, and ensure that AI-supported learning and assessment initiatives align with educational goals and principles.

5.2 Best Practices for Explainability

The goal of explainability goes beyond simply offering a clear summary of an AI system's functionality. It plays a crucial role in educational and workplace settings, particularly for the measurement of human capabilities. Providing explanations to stakeholders, including learners, teachers, administrators, and policymakers, fosters understanding and trust in AI decisions. Khosravi et al. (2022) offers four potential benefits of explainability in AI:

- **Agency:** Explainability facilitates conversations among stakeholders, enabling co-design

and co-creation. It empowers stakeholders to make informed decisions about adopting and using AI.

- **Learning Interactions:** Explainability supports the socio-cultural process of learning, where interactions between teachers and students in academic settings, and between mentors (or supervisors) and learners in the workplace, are fundamental. It can prompt dialogue on AI decisions and their implications, fostering a learning community.
- **AI Literacy:** Explainability aids in developing AI literacy, a set of competencies that enable individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI, and use AI as a tool.
- **Accountability and Trust:** Explainability helps keep educational entities and service providers accountable, addressing trust issues around the use of AI.

Explainability can be a catalyst for several societal benefits and desirable futures of human progress. It is important to consider who the explanations are for, what the purposes are, and how to effectively communicate the explanations to different stakeholders. Effective communications about AI use could come in the form of user-friendly tutorials, illustrative case studies, frequently asked questions, or other resources. While it might be challenging to simplify complex AI systems in plain language, a commitment to explainability will go a long way in fostering trust and promoting responsible AI use.

Phillips et al. (2021) outlines four core principles for achieving explainability in AI systems, which provides a framework for implementing the benefits of explainability.

- **Explanation:** For a system to be explainable, it should provide accompanying evidence, reasoning, or support for its outputs or decisions. While the accuracy or meaningfulness of the explanation is not considered here, the system needs to offer some form of explanation. For instance, consider an AI system that evaluates interpersonal communication skills during a video interview. An accompanying explanation might describe how the algorithm analyzed facial expressions, speech patterns, and eye contact to assess skills such as empathy, clarity, and engagement. This constitutes an explanation, although it may not be particularly meaningful for the candidate without additional context or information.

- **Meaningful:** AI systems should generate explanations that are understandable to their intended audience. What is considered "meaningful" will vary according to that audience's needs and expertise. Creating meaningful explanations involves understanding changing contexts and adapting explanations to meet the needs of various stakeholders, something Zapata-Rivera & Arslan (2024) call "explainable to the end user." For instance, in the AI assessment of interpersonal communication skills during a video interview, a meaningful explanation could translate the system's analysis of facial expressions and speech patterns into everyday language, highlighting how these elements indicate empathy or engagement. While one explanation might be meaningful for candidates seeking feedback, a different explanation might be more relevant for hiring managers making decisions. It is essential to tailor explanations to different contexts and stakeholders.
- **Explanation Accuracy:** This principle goes beyond the system merely producing explanations; it requires that those explanations accurately reflect how the system arrived at its output. This differs from the accuracy of the algorithm itself, instead it focuses on the explanation's accuracy. For example, in AI scoring explanation accuracy is not about the correctness of the scoring itself but about whether the provided explanation truthfully represents the behind-the-scenes operations leading to that score. If for instance, the system explains the scoring is based on the presence of keywords, but in reality, it also factors spelling, grammar, and length of the response, then the explanation is inaccurate.
- **Knowledge Limits:** Transparent and explainable AI systems should recognize and declare their knowledge limits—situations where they are not designed to operate or where their answers may not be reliable. Identifying knowledge limits can protect users from misleading or hazardous interpretations, thereby improving trust in the system.

  Building further on the AI scoring example, suppose the algorithm was trained on native English speakers' responses. Its accuracy and reliability may suffer when deployed to score responses from non-native English speakers. Explanations about the AI scoring should acknowledge this limitation. In such instances, the system should declare that it might not provide reliable scoring or feedback for non-native English speakers due to the misaligned training data.

By identifying and communicating these knowledge limits, the system can prevent potential misunderstandings or misuse of its outputs. It also contributes to building user trust, as it exhibits transparency about its operational boundaries and the reliability of its results. This open acknowledgment of the system's limitations is a crucial step toward transparent and explainable AI systems.

In short, to be considered transparent and explainable, an AI system needs to provide evidence or reasons for its outputs and actions, which should be understandable, accurate, and aware of the system's own limitations.

Best practices for achieving explainability in AI algorithms for measurement and learning include the following:

- **Explain Model Outputs in Context:** The outputs of an AI system should be explained in terms that make sense in the specific educational context in which they are used. For example, if a difficulty prediction model assigns a high difficulty level to a test item, it should clarify which features of the item (like complex language or abstract concepts) contributed to that difficulty level.
- **Highlight Influential Features:** For the outputs that the AI system generates, highlight the most influential features or factors that led to that output. This can help users understand the reasoning behind a particular decision or recommendation. For example, saliency and related measures (Arras et al., 2016; Ding & Koehn, 2021; Zhu et al., 2023) can help to highlight key facial expressions or speech patterns in a video interview that were critical to the AI system's assessment of interpersonal communication skills. If the system rated a candidate lower due to a lack of eye contact or unclear speech, the explanation should specify these aspects and clarify why they were deemed significant. Highlighting these influential features not only gives clarity on the AI's decision but also provides actionable feedback for the learner or stakeholder.
- **Disclose Limitations:** Clearly disclose the limitations of the AI system's outputs. This could involve highlighting areas of uncertainty, situations where the model is expected to perform poorly, or explaining the potential risks associated with relying too heavily on the AI system's recommendations. For example, an AI tool designed to predict student performance might not perform as expected for students who have not been represented

in the training data. It is important to communicate these limitations upfront, specifying that the tool may not have comprehensive or sufficient data to accurately predict outcomes for all student populations. It would also be beneficial to caution users against relying solely on the AI's recommendations and instead, consider them as one of many tools to inform their decisions. This responsible disclosure could build trust among users, prevent misuse of the AI tool, and promote better decision-making.

- **Use Visual Aids:** Using visual aids or interactive tools can help illustrate how the AI model operates. For example, a plot showing how different student traits influence a model's prediction of their test performance could help stakeholders visualize the model's reasoning. Saliency maps (Arras et al., 2016; Ding & Koehn, 2021), mentioned earlier, can highlight the words, tokens, audio patterns, or video cues (Zhu et al., 2023) that were most influential in making a decision about a written, spoken, or video-recorded response.
- **Involve Stakeholders in Validation:** Encourage the participation of various stakeholders, such as teachers, students, employees, employers, and policymakers, in the validation process of the AI system's explanations. This could involve conducting focus groups or interviews with these stakeholders to assess their understanding and perception of the AI's explanations. For example, do they find the explanations helpful in understanding the AI's decision-making process? Are the explanations meaningful and relatable in their context? Can they trust the AI's assessments based on the provided explanations? These insights can greatly contribute to making the AI system more transparent, beneficial, and user-friendly in measurement settings.
- **Continuous Refinement of Explanations:** Commit to persistently refining, updating, and improving the AI system's explanations based on stakeholder feedback and the latest AI research findings. This could involve revising complex technical jargon into more understandable language for students or providing more detailed explanations to satisfy the needs of teachers and administrators. With research constantly progressing, new methods of explainability may emerge that could offer better clarity and insight into the workings of the AI system. A dedication to continuous refinement ensures that the AI system remains understandable, relevant, and meaningful to all education stakeholders.

By following these principles and practices measurement organizations using AI can ensure that their AI tools are not just capable of making sound decisions but can also effectively communicate the reasoning behind these decisions. This level of transparency and explainability can cultivate a more trusting and informed environment, enabling learners, educators, and others to better understand, interact with, and benefit from AI.

5.3 Best Practices for Accountability

Accountability is critical for the ethical and responsible use of AI in education. It is about being responsible for the AI system's actions, decisions, and impacts. The stakeholders in AI, typically the developers, operators, and users of the system, should be willing to take responsibility for the outcomes produced by the AI system. They should also have processes in place to respond to any unanticipated results or detrimental impacts that might arise from the AI system's operation. Here are some best practices to foster accountability in AI measurement applications:

- **Clearly Define Roles and Responsibilities:** Clearly articulate the roles and responsibilities of all stakeholders involved in the development, deployment, and use of the AI system, including developers, educators, administrators, and learners associated with the creation, deployment, and use of the AI measurement system. This practice incorporates responsibilities for the collection and protection of learner data, AI model development and updates that align with the learning and measurement goals, evaluation of the AI system's performance, making decisions based on the AI system's results, and addressing any errors or unanticipated outputs from the system.
- **Establish Accountability Mechanisms:** Establish mechanisms that allow stakeholders to hold the AI system (and its operators) accountable. This could include grading and classification systems for AI systems' behaviors, standard operating procedures for remediation in case of errors or harmful impacts, and mechanisms for dispute resolution or compensation for harm. For example, schools could develop a standardized procedure for teachers to dispute and correct an AI system's outputs. Additionally, a grievance system could be established, allowing learners or test takers to report significant concerns with the AI outputs. If the AI system caused harm to a student's academic record, corrective measures such as adjusting the student's scores or providing additional

educational support would be implemented. These structures maintain accountability for the use of AI in educational and assessment applications.

- **Publish Regular Audit Reports:** Regularly publish audit reports detailing the AI system's performance, the extent of alignment with ethical and fairness principles, the actions taken to improve the system, and any issues or instances of harm that arose. These reports can ensure transparency about the system's operation and the developers' or operators' efforts to hold themselves accountable. For example, an audit report for an AI-generated assessment might include data on how accurately and consistently the AI system scores tests compared to human-generated assessments, the steps taken to ensure that student data is kept private and secure, and any biases that may have been detected in its scoring patterns. It could also outline any issues that arose during the assessment process, such as technical glitches or inappropriate score adjustments, along with what was done to address these issues and prevent their recurrence.
- **Engage in Open Dialogue with Stakeholders:** Encourage open dialogue with stakeholders about the AI system's impacts and the developers' or operators' accountability efforts. This could include, for example, holding regular meetings with teachers, administrators, and students, soliciting public feedback on the AI system's use, or involving external committees in reviewing and guiding the system's operation.

Accountability in AI measurement applications is not only about providing explanations for the AI system's operations. It is also about making a commitment to respond promptly to any potential issues such as unfair grades, biased performance evaluations, or adverse effects on learning processes or career advancement. It cultivates trust among learners, educators, employees, and policy makers and ensures its utilization aligns with the core values of education and the specific goals of the institution or course.

6 The Principles and Practices of Impact and Integrity

The principles guiding the use of AI for measurement, specifically focusing on impact and integrity, revolve around aligning with the learning objectives and values. These principles are essential to promote the valid use of information provided about learners' knowledge, skills, and abilities. By being attentive to the potential negative implications and respecting personal autonomy, AI can promote human progress without compromising integrity.

Principles related to impact and integrity include the following:

- **Alignment with Learning Objectives:** The application of AI in for measurement should support the learning objectives including the curriculum, specific learning outcomes, and personal growth and development. The AI system needs to provide accurate predictions and recommendations on learners' behaviors, skills, and progress in alignment with these goals.
- **Validity:** The inferences made by AI should be valid for their intended use. The inferences should accurately reflect the learner's knowledge, skills, and abilities in the domain being assessed or measured.
- **Mitigating Negative Consequences:** The potential negative effects of AI, such as over-reliance on technology, reduced teacher-student interaction, or the risk of privacy infringement, need to be actively considered and mitigated.
- **Respect Personal Autonomy:** The use of AI in learning and measurement must respect the autonomy and individuality of all learners. Algorithms should be designed and implemented in a way that respects individuals' rights to make choices about their education.

The use of AI for measurement must uphold the principles that respect individual rights, promote human progress, and emphasize educational integrity. This necessitates an ongoing commitment from educators, developers, and policymakers to ensure that the use of AI in education serves as a tool to support human progress rather than a replacement for human roles.

6.1 Best Practices for Aligning with Learning Objectives

The deployment of AI for learning and measurement should fit into a broader strategy for improving student learning and supporting human progress. Education providers and designers must clearly define their goals for AI integration, such as enhancing personalized learning, facilitating adaptive instruction, supporting skill development, or improving administrative efficiency. Aligning AI usage with these objectives ensures the selection and design of systems that possess features and capabilities aligned with these goals.

AI usage should align with the organization's context, the learning environment, and the diverse needs of learners. For example, if the goal is to support a wide range of learning paths,

AI tools should offer features that accommodate varied educational approaches and abilities. Additionally, AI should align with pedagogical principles, promoting deep learning, critical thinking, and collaboration. This practice ensures the AI tools enhance the educational process in line with the institution's priorities.

Finally, institutions should establish clear metrics for measuring the effectiveness of AI deployment. These metrics can include improvements in learning outcomes, increased learning engagement, more personalized learning experiences or any other such indicators. The data gathered can be used to continually refine and optimize the AI system in line with the learning goals.

Some practices to help achieve the alignment of the AI with the learning goals include the following:

- **Clarify Purpose and Objectives:** Before developing or selecting an AI tool for assessment, it is essential to define the purpose of the assessment (e.g., formative, summative, diagnostic), the learning objectives it should measure, and how its results will be used. As Messick (1992) states, "A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society." This step helps to define the student, or competency model in the Evidence Centered Design (ECD; Mislevy et al., 2003).
- **Evaluate Appropriateness of AI Algorithms:** The chosen AI algorithms need to fit the educational goals of the assessment. Not all algorithms are suited for all types of assessments or domains. For example, if the goal of an assessment tool is to evaluate a student's problem-solving skills, the underlying AI algorithm should have capabilities to analyze complex cognitive behaviors beyond simple recall of information. In conversation-based assessments of complex skills, generative AI should be capable of guiding discussions to elicit evidence of skill mastery or proficiency, prompting responses that demonstrate critical thinking and adaptive reasoning.
- **Consider the Educational Context:** The context in which the AI system will be used should be extensively considered in its conception. It includes the socioeconomic, cultural and language background of students and the infrastructure of the institution, its

pedagogical approach, and general comfort with technology. AI tools need to be flexible enough to be applied in different situations or be designed for specific contexts.

- **Align With Current Systems and Processes:** The implementation of AI for learning and measurement should complement and enhance existing programs and services. For example, AI-based assessment tools should work with existing learning management systems to provide a seamless integration for learners, educators and administrators when applied in such settings. Similarly, the insights generated by AI analyses should be directly actionable within the existing pedagogy.
- **Involve Stakeholders:** To align AI with learning goals, engage different stakeholders—educators, administrators, learners, parents, technology providers, and policymakers—in the design, implementation, and evaluation process. This collaborative approach ensures that AI tools are tailored to the needs and expectations of end-users and align with broader learning objectives.
- **Employ Human-in-the-Loop Approaches:** The use of AI for measurement and learning should maintain a strong human presence in the decision-making process. As AI is increasingly being used to assess student performance and recommend learning materials, the ultimate authority over these judgments should reside with human educators. For example, if an AI tool determines a student has difficulty in understanding specific mathematical concepts, the decision to modify the learning path should be overseen or even finalized by the teacher. This approach, often termed as a "human-in-the-loop" strategy, respects the professional insights of educators, ensuring that AI operates as a supportive tool, not as a substitute for their critical role.
- **Encourage Social Interaction and Collaboration:** AI tools should be designed to foster a sense of community and collective learning among students. For instance, AI-powered group learning platforms could be utilized to create virtual collaborative projects, offer real-time brainstorming sessions, or facilitate peer-to-peer discussions. These tools should actively promote, not impede, the valuable interactions that students have with their classmates and teachers. While AI may provide personalized learning experiences, it is crucial not to isolate students or inadvertently diminish the significance of communal learning and human connection in education.

- **Address Accessibility:** AI applications used in learning and assessment platforms should be designed with a focus on universal access. This involves being mindful about the diversity of students, who come from varied cultural backgrounds, speak different languages, have unique learning abilities, and follow distinct learning paths. For instance, AI-powered measurement platforms could include multi-language support, features for visually or hearing-impaired learners, or distinct learning paths based on a learner's preferences. Access to technological advancements like AI in education must not create or widen the disparity among students but should rather work toward supporting human progress and socioeconomic mobility for all learners.

The goal of integrating AI into measurement and learning applications should go beyond simply automating tasks or improving efficiencies. It should be about improving the learning experience, maximizing human outcomes, and supporting the development of all learners. By aligning AI use with learning goals and objectives, organizations can leverage AI to support human progress. However, this alignment is not a passive or one-time process, it requires ongoing engagement and monitoring from all stakeholders.

6.2 Best Practices for Validity

In addition to following the standards for validity laid out in AERA et al. (2014), the use of AI in learning and measurement systems, the validity of the AI system's output can be improved by considering the following practices:

- **Adopt an Argument-Based Approach:** An argument-based approach to validation entails constructing and critically evaluating arguments for and against the intended interpretations and uses of the AI system's score outcomes. This encourages comprehensive thinking about the potential validity threats and the evidence needed to counter them. In an educational context, this focuses on the academic skills, knowledge, attitudes, and competencies that the AI system is designed to assess. By defining intended interpretations and uses, validity threats and counterarguments can be clearly identified, assessed, and mitigated.
- **Ground the AI System in Learning Science:** The function of AI in learning and assessment should be firmly grounded in accepted learning theories. Such theories should inform the construction and application of AI, becoming the foundation for its validation

and effectiveness. For example, by incorporating theories like 'zone of proximal development' (Vygotsky, 1978), AI tools could be designed to pinpoint the gap between what a student already knows and what they are ready to learn next, facilitating personalized assessments that guide students along their learning path.

- **Compare AI Output to Multiple Sources of Evidence:** Use multiple sources of data in determining the validity of AI outcomes. These sources can include observational data, self-reports, peer reports, and any other relevant information. For example, in an audiovisual-based interpersonal skills assessment, comparing AI-generated inferences or scores with peer reviews and human analysis of video can help validate the system's evaluation of communication skills. Comparing the output of AI systems to multiple sources of data can provide evidence for the validity argument.
- **Evaluate Accuracy, Reliability, and Fairness:** Establish precise performance metrics that define acceptable levels of accuracy, fairness, and reliability. Monitoring and maintaining these metrics will ensure that validity remains a priority in the AI system. For example, McCaffrey et al. (2021) provides guidance on accuracy and fairness metrics that can be used to evaluate automated scores and provide evidence to support their valid use in assessment.

Following these practices can help promote the effective use of AI for learning and measurement, providing insights about the intended learning objectives and yielding meaningful, trustworthy results.

6.3 Best Practices for Mitigating Negative Consequences

Considering the principles above, the following best practices can be employed to further mitigate the potential negative consequences of using AI for measurement and learning purposes.

- **Evaluate and Improve Algorithmic Accuracy:** AI system's accuracy should be regularly evaluated to avoid wrong judgments or outcomes. If inconsistencies or errors are found, they should be promptly addressed. Ensuring the accuracy of the AI's algorithms will not only improve the system's overall performance but will also establish trust among the users.
- **Promote Data Literacy Among Stakeholders:** Stakeholders, such as educators,

learners, employers, and industry experts should be provided with adequate training and resources to understand how AI systems use and interpret data, helping them make informed decisions about using these technologies within the lifelong learning journey.

- **Engage Stakeholders Continuously:** Encourage active collaboration among stakeholders such as educators, learners, employers, and industry experts in the adoption, implementation, and refinement of AI systems for the measurement of human capabilities. Their input is invaluable in iteratively enhancing these systems, ensuring they effectively support the lifelong learning journey.
- **Personalize within Boundaries:** AI can enhance personalization in education and skills development, but it is crucial to avoid creating environments where learners are exposed to limited information or perspectives. Striking the right balance between tailoring experiences and offering diverse learning opportunities ensures a well-rounded development.
- **Implement Ethical AI Training Programs:** Develop training programs that not only provide technical understanding of AI systems but also instill ethical considerations in their design, implementation, and use.
- **Emergency Stop Mechanism:** Implement a "stop button"—a mechanism that allows human operators to immediately interrupt or halt AI system operations in the event of unintended, unsafe, or harmful behaviors, as recommended in international AI ethics guidelines.

These best practices, while not exhaustive, offer a roadmap to mitigating potential negative consequences of implementing AI for measurement. It is essential to note that regular monitoring and updating of these practices is crucial to align with the evolving AI technologies and the ever-changing human developmental landscape.

6.4 Best Practices for Respecting Personal Autonomy

Personal autonomy refers to a person's ability to make choices and decisions based on one's own principles and values. In the context of AI systems, respecting personal autonomy involves putting mechanisms into place that allow individuals to understand how the systems work and affect their lives, and to exercise personal control over their engagement with these

systems.

- **Respect Individual Choices:** Users should have the opportunity to control how their data is being used. Consent mechanisms should be clear and explicit, providing stakeholders, such as learners, educators, employees, and employers with information about what the AI is doing and how it will affect them.
- **Promote Self-Directed Learning:** AI systems should be designed to act as supportive tools that encourage students to take charge of their own learning. Feedback provided by AI should be used to guide learning and development choices rather than dictate them.
- **Avoid Over-reliance on AI:** While AI can be a helpful tool in education, over-reliance on it can inhibit personal autonomy. Learners should still have opportunities for human interaction and guidance.

These practices ensure that while using AI technologies for learning and measurement, the individual's power to make decisions and exercise control is retained, thereby promoting a more engaging, participatory, and empowering experience.

7 The Principles and Practices for Continuous Improvement

Continuous improvement in AI applications is essential to ensure that AI systems remain effective, efficient, fair, and aligned with learning and development goals. Implementing practices for continuous improvement allows organizations to adapt to changing circumstances, address emerging challenges, and enhance the overall quality of their AI systems. Principles for achieving continuous improvement in AI applications for measurement include the following:

- **Adaptability:** AI systems should be designed to be flexible and adaptable, allowing for updates and modifications in response to changing developmental needs or technological advancements.
- **Learning:** Continuous improvement involves learning from past experiences, feedback, and data to enhance the performance of AI systems over time. This learning process should be intentional and systematic.
- **Data-Driven:** Continuous improvement should be data-driven, relying on evidence and feedback to identify areas for enhancement and measure the impact of changes made to

the AI system.

- **Feedback Loops:** Feedback from all stakeholders, including students, should be actively sought and incorporated into the AI system. This feedback can provide valuable insights into how the AI is being used, its impact on learning outcomes, and potential areas of improvement.

Through the implementation of these practices for continuous improvement, we can ensure the functionality, efficacy, and positive impacts of AI in the field of educational testing.

7.1 Best Practices for Continuous Improvement

Continuous improvement is essential to ensure the effectiveness and relevance of AI applications in educational testing. Practices to promote the continuous improvement of AI applications in education include the following:

- **Modular Design:** Implement a modular design approach that allows components of the AI system to be updated, replaced, or enhanced independently. For instance, in a conversation-based assessment of complex skills, the generative AI responsible for guiding the conversation can be updated independently from the algorithms used for evidence extraction and scoring. This approach enables incorporating new conversational techniques without altering the core assessment algorithms, maintaining flexibility and adaptability.
- **Scalable Design:** Design AI systems to be scalable, allowing them to handle increased data volume, user load, and additional functionalities without compromising performance. Scalability prepares the system for growth and ensures usability in diverse learning settings.
- **Conduct Regular Evaluations:** Conduct regular assessments and evaluations of AI systems to measure their performance, reliability, validity, fairness, and overall impact on learning outcomes. Use evaluation results to identify areas for improvement.
- **Implement Feedback Loops:** Establish feedback mechanisms that allow users, such as employees, employers, learners, and educators to provide input on the usability, effectiveness, and relevance of the AI system. Use feedback to drive continuous improvements.

Following these practices for continuous improvement can help ensure that the application of AI systems for measurement and learning is effective and aligned with the evolving needs of stakeholders. Continuous improvement can help foster innovation, promote progress, and strengthen the impact of AI technology to support human progress.

8 Concluding Remarks

At ETS, we recognize the potential of AI to contribute to the science of measurement and learning more broadly. We also recognize the challenges it poses. Our commitment to responsible AI guides how we incorporate this technology into our work. Through rigorous research and development, thorough quality assurance, stakeholder consultation, continuous improvement, transparent communication, and collaborative governance, we strive to apply AI in a manner that benefits all learners and stakeholders. Ultimately, our aim is to ensure that AI enhances the validity, reliability, fairness, and utility of our measures as they support human progress.

It is important to acknowledge that the best practices and principles presented in this document are not exhaustive; as AI technologies and educational needs continue to evolve, so too must our approaches. We encourage continued reflection, learning, and adaptation to emerging developments, so that our commitment to responsible AI remains strong and relevant.

Appendix

Existing Guidance on AI and Educational Testing

Organizations such as the National Institute of Standards and Technology (NIST), the US Department of Education (US DOE), the Organisation for Economic Cooperation and Development (OECD), The European Commission, and UNESCO have proposed principles and guidelines to address these challenges and risks. These organizations' guidelines provide recommendations and best practices for the development and use of AI in education and in general. Those that focus on the application of AI in education emphasize the need to protect and promote human values and interests in the educational process. Other organizations such as the American Psychological Association (APA), the American Education Research Association (AERA), the National Council on Measurement in Education (NCME), and the International Test Commission (ITC) have similarly developed broad principles and standards for the design,

development, and use of educational assessments. This section comprises summaries of guidance from leading organizations in the field, beginning with NIST.

National Institute of Standards and Technology

The Artificial Intelligence Risk Management Framework developed by NIST (2023; see also Schwartz et al., 2022) outlines several essential principles that govern the trustworthiness of AI systems. These principles include validity, reliability, safety, security and resilience, accountability, transparency, interpretability, and privacy. NIST discusses the criticality of balancing these characteristics based on the context in which an AI system operates to create trustworthy AI.

However, addressing each principle individually does not ensure complete trustworthiness since trade-offs often exist between these characteristics depending on different contexts or situations. For instance, managing risks may involve making decisions between optimizing for explainability versus achieving privacy, or balancing predictive accuracy against ensuring fairness. The framework also emphasizes that transparency is crucial for accountability while explainability and interpretation are necessary for promoting the understanding of system outputs by all stakeholders.

In addition, NIST recommends considering cultural and contextual differences regarding the perception of fairness and trustworthiness among various stakeholders within the AI lifecycle such as developers versus deployers, designers versus users, and the differences across demographic groups.

Based on these principles NIST provides a structure for understanding, managing, and mitigating risks associated with AI systems. The framework is composed of four main functions:

- **Govern:** Governance is a cross-cutting function designed to be considered throughout the AI lifecycle. It involves creating a culture of risk management within the organizations involved in any stage of an AI system's lifecycle. It includes creating policies for identifying potential impacts and developing mechanisms to handle them.
- **Map:** Mapping helps establish context to frame risks related to an AI system through activities like understanding intended purposes or objectives, anticipating possible negative impacts, recognizing when systems are functional or non-functional in their intended context, and others.

- **Measure:** The measurement function uses quantitative and qualitative methods to analyze and monitor risks identified in the mapping stage. This measurement function informs later decisions in the management phase.
- **Manage:** This entails allocating the resources that are necessary to handle the mapped and measured risks regularly, following the practices and guidelines defined in the governance stage.

United States Department of Education

The U.S. Department of Education's (DOE) publication, *Artificial Intelligence and the Future of Teaching and Learning* (Office of Educational Technology, 2023), recognizes the potential of AI to transform education. It suggests that AI could facilitate new avenues for interaction between students and teachers, adjust to individual student differences in learning, improve adaptivity based on a student's real-time learning processes rather than simply whether or not an answer is correct, augment feedback mechanisms provided to both educators and learners, all while potentially supporting educators, administrators and policy makers.

The U.S. DOE guidance also strongly advocates for human involvement in the educational processes that rely on AI. The publication offers several recommendations about how AI can be implemented in educational systems in responsible ways. These recommendations include ensuring human oversight ("humans-in-the-loop"), alignment with a common vision for education, application design using contemporary pedagogical principles (e.g., inclusive design), fostering trust through transparency in technological applications among users, along with an emphasis on context-based research & development. The DOE guidance concludes with calls for clear data privacy regulations specific to educational technology applications incorporating the use of AI.

UNESCO

UNESCO's *Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 2022; see also, UNESCO, 2023) provides an exhaustive set of guidelines for using AI technology in education. It highlights that while AI has numerous advantages, it is not necessarily designed to guarantee human and environmental welfare. It suggests that any application of AI in an educational setting must be aligned to legitimate educational objectives and should not cause harm. The UNESCO report further recommends that the decision to utilize AI should adhere to

three main principles: suitability for a specific objective; respect for fundamental values and human rights; and a sound scientific basis.

The UNESCO document stresses the importance of fairness, advocating against discrimination in access to AI and technology across age groups, cultural backgrounds, linguistic communities, etc. The authors further argue that risks resulting from unintended consequences or security vulnerabilities require rigorous oversight throughout the AI lifecycle. Transparency also plays a vital role; individuals impacted by algorithmic decisions should have access to information explaining how these decisions were made. Furthermore, activities related to the collection and use of data must strictly follow international law and other established legal frameworks.

UNESCO emphasizes that AI education is a crucial element in raising public awareness about the implications of AI use, particularly its potential impacts (both positive and negative) on society. The documents emphasize the need to foster literacy and skills development to promote the effective and responsible use of AI. Lastly, UNESCO urges multi-stakeholder participation in order to help ensure that the benefits of AI are equitably shared with all individuals. This includes creating regulations that are inclusive and adaptable to technological changes.

OECD

OECD proposed a series of guidelines for the responsible development and use of AI (OECD, 2019a, 2019b). The OECD guidance is intended to apply broadly to all stakeholders involved in AI. These stakeholders include developers, users, and regulators. In the case of education, this would include students, teachers, administrators, and policy makers. The document emphasizes that their guidelines should be viewed as connected, all contributing toward the overall goal of responsible AI use.

Inclusive growth is one of the key principles of these recommendations. The principle advocates that AI technologies should be developed with an aim to provide benefits for all people and the planet by supporting human progress, promoting inclusivity, reducing socioeconomic disparities, and to protect the environment.

A significant emphasis of the OECD recommendations is on human-centered values, including respecting legal norms and democratic values like freedom and privacy. Implementing appropriate AI practices consistent with current state-of-the-art technology and policies can help to adhere to these principles.

Transparency forms another critical aspect within the OECD recommendations. The OECD report emphasizes that the individuals interacting with the AI system should understand how it works. Stakeholders must be made aware of their interactions with these systems, including potential adverse effects, which can help them challenge any questionable outcomes.

The robustness, security and safety of AI systems is another key area addressed by the OECD recommendations. The recommendations urge the development of AI systems that function correctly under normal and predictable adverse conditions, without posing unreasonable risks. To ensure this, mechanisms for traceability should be put in place along with systematic risk management approaches to address potential concerns including privacy violations, digital security breaches and bias.

Finally, the OECD places a strong emphasis on accountability. All parties involved in developing or using an AI system must be accountable for its proper functioning while also respecting all other principles of responsible AI use.

The European Commission's Assessment List for Trustworthy AI

The European Commission's High-Level Expert Group on AI has developed ethics guidelines (High-Level Expert Group on AI, 2019) and an *Assessment List for Trustworthy AI* (ALTAI; High-Level Expert Group on Artificial Intelligence, 2020) for ensuring that AI systems are designed and used responsibly. A key aspect of ethics guidelines and the ALTAI is the Fundamental Rights Impact Analysis (FRIA) that evaluates potential negative discrimination by AI systems across a range of sensitive attributes including race, gender, religion, political view, disability, or age, among other factors. The AI systems are also assessed from the perspective of child protection, data protections in line with European Union's General Data Protection Regulation (GDPR), and respect for freedom of expression, information, and assembly.

One of the essential requirements of ALTAI is human agency and oversight to ensure respect for human autonomy and decision-making. This means AI systems should uphold fundamental rights while enabling a democratic, flourishing society. Similarly, technical robustness and safety are stressed, ensuring AI systems are reliable, resilient, and deliver trustworthy services. They should be developed with proactive approaches to risk and function as intended, minimizing potential harm.

Privacy and data governance are other integral parts of the ALTAI; AI providers should provide adequate protection to personal data used by AI systems and ensure the quality and

relevance of the data for the application. The requirement of transparency in AI involves traceability, explainability, and open communication about the system's limitations, which in turn should increase users' trust in the application.

The concept of diversity, non-discrimination, and fairness explores AI systems' capability to promote inclusion and diversity throughout the system's lifecycle and to eliminate any biases. Attention is also given to societal and environmental well-being, emphasizing that the broader society and the environment should be regarded as stakeholders in AI's developmental and deployment phases.

Lastly, accountability requires suitable mechanisms to ensure responsibility in developing, deploying, and using AI systems, with transparent risk management and guidelines for redress when adverse impacts occur. The ALTAI suggests rigorous standards to ensure an AI systems' trustworthiness while fostering their benefits across society.

The Joint Standards for Educational and Psychological Testing

The *Joint Standards for Educational and Psychological Testing*, produced by AERA, APA, and NCME (2014), provide comprehensive guidelines for test creation, execution, and analysis. These standards emphasize validity (tests measuring what they are designed to assess), reliability (consistency of test scores), and fairness (accessibility and potential bias).

The Standards demand transparency in every aspect of the test's development and administration. The Standards recommend that testing organizations publish technical manuals to document the test development process, provide evidence of the test's validity, and offer guidelines for interpreting scores and testing procedures. The Standards also discourage the misuse of the tests for purposes not supported by the technical documentation.

The Standards stress the importance of evidence-based decisions, requiring that test developers and administrators continually gather and analyze data to improve the validity and reliability of their tests. Lastly, the Standards outline the responsibilities of the test user, who is expected to follow ethical testing practices, familiarize themselves with the test content, and make informed decisions based on the test results.

The Standards, which are now 10 years old, provide a foundational framework for the development and usage of tests, but do not explicitly address the intricacies of modern AI and generative AI in educational testing. However, the Standards can be used as a broad starting point for developing principles and best practices for the use of AI in educational measurement.

International Test Commission (ITC)

The *International Test Commission and Association of Test Publishers' Guidelines for Technology-Based Assessment* provide a detailed framework for implementing technology-based tests. The guidelines highlight the need for clear objectives, valid and reliable assessments, and the use of universal test design principles to accommodate diverse test takers, including those with special educational needs. A particular emphasis of the TBA guidelines is that the test design and content should align with the test's purpose and that contingency plans should be put in place to handle potential technological disruptions.

The TBA guidelines emphasize that data governance is crucial in the digital age; discoverability, availability, and quality of data are vital for assessment development and administration. The guidelines discuss the need for security measures to promote data confidentiality and privacy and ensure compliance with legal and professional standards. The integration of assessment data with other educational systems is also recommended to support comprehensive analysis and seamless data integration across systems.

The guidelines stress the importance of evaluating psychometric and technical quality to ensure accurate and valid measurements. Attention is needed for validation processes, score comparability, score equating, and measurement quality. Equity, fairness, and accessibility are also crucial, with universal test design principles helping to accommodate a wide variety of test takers. Potential biases should be addressed in the design and administration of the test.

Lastly, the guidelines emphasize the importance of test security, recommending written security plans and ongoing technology evaluations. Preventive measures against fraud, enhanced detection techniques, and adequate response measures are also advised. Compliance with personal data regulations, transparent communication of privacy policies, risk assessments, and the application of privacy by design principles are also recommended. The guidelines aim to provide a balanced approach to technology-based assessments, by considering technical advances in concert with privacy, accessibility, fairness, and data security issues.

A Synthesis of the Existing Guidance

Synthesizing the existing guidance from the Department of Education, UNESCO, OECD, AERA, and the ITC, we can identify several core principles that should guide the responsible use of AI in educational testing:

- **Alignment with Educational Values and Goals:** AI applications should align with broader educational values and goals. This involves using AI to enhance learning and teaching, rather than focusing solely on efficiency or cost-cutting.
- **Validity and Reliability:** AI systems, particularly in educational testing, should yield valid and reliable results. This not only involves accurately and consistently measuring the intended educational outcomes but also ensuring the representation of the underlying constructs accurately. The AI should take into account the complexity of learning objectives, skills, knowledge, or attributes that it attempts to measure, to minimize mismatches between the intended and assessed constructs.
- **Fairness and Non-Discrimination:** AI applications should promote equity by assuring fair and non-discriminatory treatment of all individuals. This includes elements of accessibility but also needs to encompass an equitable approach in the application and effects of the AI system. Fairness should be upheld in processes and outcomes, ensuring that no particular group is unfairly advantaged or disadvantaged by AI decisions or results. This involves acknowledging and addressing potential biases in AI algorithms, ensuring that AI responses do not perpetuate existing inequalities, and making certain that all students, irrespective of their socioeconomic, cultural, and linguistic backgrounds, have equal opportunity to benefit from the AI application.
- **Explainability, Transparency and Accountability:** Explainability and transparency in AI systems refer to the comprehension and the communication of how these systems function, make decisions, and generate results. Particularly in educational testing, it is integral that AI's processes and results align with the educational goals or constructs being measured. Stakeholders should be able to validate that the AI's outcomes are consistent with the intended educational objectives. Failure to provide this explanation might lead to a lack of trust in the system. Moreover, accountability is a key requirement. There should be mechanisms in place for stakeholders to question, challenge, or appeal the decisions made by the AI. This transparency, explainability, and accountability contribute to building a more reliable and effective educational AI system.
- **Privacy and Security:** AI in education should prioritize data privacy and security. This includes protection against data breaches and ensuring that personal information is

handled in accordance with legal and ethical guidelines.

- **Human-Centered Approach:** AI should be developed and used with the primary goal of benefiting and promoting the interests and well-being of all stakeholders including students, educators, and parents. While technological advancement and efficiency are important, they should not overshadow the importance of serving the users' needs, cultivating their abilities, and preserving their values. To ensure this, a 'human-in-the-loop' approach should be maintained where educators, students, and other stakeholders collaborate in developing, controlling, and implementing AI systems, ensuring their alignment with human values and educational goals.
- **Inclusivity and Accessibility:** AI in education should be inclusive and accessible to all, regardless of their individual abilities or needs. This can involve applying universal design principles to accommodate diverse learners.
- **Ongoing Monitoring and Improvement**: The performance of AI systems in education should be continuously monitored and improved. This includes regular evaluations of the validity, reliability, fairness, and other characteristics of the AI system.
- **Ethical Use:** AI should be used ethically, caring for the interests of all stakeholders, respecting the rights of all individuals involved, and abiding by established ethical guidelines and standards.

Glossary

This glossary defines terms found in the main text, along with additional related concepts that provide helpful background for understanding AI and measurement.

algorithm. An algorithm is a step-by-step procedure or a set of rules for solving a particular problem or completing a specific task.

AI lifecycle. The stages involved in the development, implementation, and maintenance of artificial intelligence systems, typically including data collection, model training, model evaluation, deployment, and monitoring.

artificial intelligence (AI). A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content,

recommendations, or decisions that can influence physical or virtual environments (Russell et al., 2023).

automated item generation. Automated item generation refers to the use of AI and computational means to generate test items for educational assessments.

automated scoring algorithm. An automated scoring algorithm is a computational procedure used in educational testing to predict or determine scores for test items or responses automatically. These algorithms typically use natural language processing, and statistical or machine learning techniques to generate predicted scores based on patterns or associations found in the data.

bias (statistical). Statistical bias refers to the tendency of a statistical measure or estimator to systematically overestimate or underestimate the true value of a parameter being estimated. Statistical bias can lead to unfair or inaccurate outcomes, particularly when evaluating student performance or making decisions based on AI-generated insights.

bias mitigation. Bias mitigation refers to the process of reducing or eliminating biases in AI systems to ensure fair and equitable outcomes for all individuals.

cost-sensitive learning. Cost-sensitive learning is a method in machine learning aimed at minimizing the misclassification costs, as opposed to merely minimizing typical error measures like mean squared errors. In cost-sensitive learning, different types of errors are assigned different costs (e.g., positive errors for one demographic group might be considered more costly), and the training algorithm works to reduce the total cost.

counterfactual fairness. Counterfactual fairness is a concept in the field of AI and machine learning that pertains to ensuring that the decisions made by an AI system remain consistent even if certain features or attributes of the individuals involved were hypothetically changed.

deep learning. A subset of machine learning involving neural networks with many layers to analyze complex data patterns.

disparate impact. Disparate impact refers to practices that adversely affect one group of people of a protected characteristic more than another, even though rules applied are formally neutral. Although the underlying practice is not discriminatory in its intent, it may have discriminatory effects.

disparate impact analysis. statistical methodology used to identify instances of indirect discrimination in practices or policies. It involves examining the outcomes of decisions to see if they disproportionately impact certain groups protected by anti-discrimination laws, even if the decisions were not intended to discriminate.

embedding. Embeddings, in the context of AI and machine learning, refer to the practice of transforming categorical variables or discrete data structures into continuous vectors that can be used by a model. Embeddings are used in a variety of applications but are perhaps most widely known in the context of natural language processing. Word embeddings, for example, transform words or phrases into numerical vectors, preserving semantic relationships between different words.

equity. Equity in education refers to the recognition and provision of differentiated resources and opportunities to cater to the unique learning needs of all students so that every individual has an equal opportunity for academic success. This notion of equity underscores that treating students fairly may sometimes mean treating students differently by accommodating to their unique circumstances, abilities, and backgrounds.

explainability. Explainability in the context of AI refers to the ability to understand and interpret the decisions or outputs produced by an AI model. This includes understanding why the model made a specific prediction, how different inputs influence the model's predictions, and the general logic that the model uses to make predictions.

fairness. Fairness, in the context of AI and educational testing, refers to the principle of ensuring that AI systems do not favor or disadvantage any particular group of individuals based on characteristics such as socioeconomic status, culture, race, gender, or ability.

fairness-aware algorithms. Fairness-aware algorithms are machine learning algorithms that are specifically designed to avoid unfair biases in their predictions. These algorithms either modify existing machine learning techniques or incorporate fairness considerations into their design to ensure their predictions do not disproportionately disadvantage or benefit any one group, especially those based on sensitive features like gender, race, or age.

fine tuning. Fine tuning, in the context of AI and machine learning, refers to the process of adjusting a pre-trained model to better adapt to a specific task. This is done by continuing the

training process on a secondary dataset that is more specific to the task, with the aim of refining the model's parameters to better align with the task at hand.

GDPR. The General Data Protection Regulation. This is a regulation in EU law that focuses on data protection and privacy in the European Union and the European Economic Area. It also addresses the transfer of personal data outside these areas. The GDPR aims to provide individuals with control over their personal data and to simplify the regulatory environment for international business.

generative artificial intelligence (GenAI). GenAI refers to AI systems that can generate novel content, such as text, images, audio, and video, based on data inputs and user preferences.

group fairness. Group fairness refers to the principle of ensuring that AI systems do not exhibit systematic biases that disadvantage certain groups of individuals based on characteristics such as race, gender, socioeconomic status, or other protected attributes.

inclusivity. Inclusivity is the practice or policy of providing equal access to opportunities and resources for people who might otherwise be excluded or marginalized, such as those having physical or mental disabilities or belonging to other minority groups.

individual fairness. Individual fairness is a principle in AI and machine learning that focuses on treating similar individuals similarly.

injection attack. Injection attacks are a type of security vulnerability where an attacker is able to inject malicious data or code into a system. This data is then processed by the system, leading to unexpected and potentially harmful results. The most common type of injection attack is a SQL injection, where malicious SQL code is inserted into a query, allowing an attacker to manipulate the database, potentially gaining access to confidential data or modifying data in unauthorized ways.

language model. A Language Model or Large Language Model (LLM) is an AI or machine learning model that is trained to understand, generate or work with human language, and predict a word, or a sequence of words, in a sentence. It does so by calculating the probability of occurrence of a specific word given a sequence of words, called the context.

learner. In an educational context, a learner refers to an individual who is engaged in the process of acquiring knowledge or skills. This term emphasizes the active role of the individual in their

own education, as opposed to more passive terms such as student or passive learner.

machine learning (ML). A branch of AI focusing on building systems that learn from data.

model training. The process of training an AI algorithm using data to learn patterns and make predictions or decisions.

model evaluation. The process of assessing the performance of an AI model by measuring its accuracy, bias, and/or other metrics.

natural language processing (NLP). AI techniques that enable machines to understand and interpret human language.

neural network. Models and algorithms modeled after the human brain, designed to recognize patterns, and interpret complex data inputs. Neural networks are fundamental to deep learning processes within AI, enabling the system to learn and make decisions from data.

pre-trained model. A pre-trained model refers to an AI or machine learning model that has already been trained on a large-scale dataset. These models are often trained on general tasks, such as image classification or language processing, and are designed to capture wide-ranging patterns in the data on which they are trained. The parameters of these models are often shared publicly and can serve as a starting point for further model development.

pseudonymize. The process of disguising identities so that they cannot be connected to their real-world identities without additional information that is held separately. This helps to protect privacy by separating data from direct identifiers so that linkage to an identity is not possible without additional information that is held separately.

redress. Redress refers to the process of addressing or resolving grievances, complaints, or disputes in a fair and equitable manner. It may involve compensating individuals for harm, providing assistance or support to address issues, or making changes to prevent similar problems from occurring in the future. Redress is important in ensuring accountability.

representative data. Representative data refers to a sample of data that accurately reflects the larger population or dataset from which it is drawn. In the context of AI, a representative dataset means that the AI system has been trained on data that includes a wide variety of situations, inputs, and variables that the AI system will encounter during its operation.

saliency measure. A saliency measure is a metric or technique used in AI and machine learning to quantify the importance or relevance of input features to the output prediction made by a model. In the context of educational testing, saliency measures can help identify which factors or characteristics in student data are most influential in determining their performance or outcomes.

secure by design. A principle that encourages the consideration and integration of security measures from the very beginning stages of product or system development.

sensitive attribute. A sensitive attribute is any characteristic or trait that should not influence the decisions or predictions of the AI system due to ethical, legal, or fairness considerations. Examples of sensitive attributes commonly considered in fairness and bias analyses include race, gender, and socioeconomic status.

sensitivity. In the context of AI and machine learning, sensitivity refers to the measure of the true positive rate. It is the ability of the model to correctly identify positive instances, i.e., the number of correct positive predictions compared to the actual number of positives. It is also known as recall, hit rate, or true positive rate.

socio-educational. The intersection of social and educational factors or processes. This term is used to describe the ways in which social conditions, contexts, and structures influence educational environments, practices, and outcomes. It encompasses aspects such as socioeconomic status, cultural background, community resources, and social dynamics, and how these elements affect the learning experiences and opportunities available to individuals.

specificity. In the context of AI and machine learning, specificity refers to the true negative rate, i.e., the ability of a model to correctly identify negative instances. It shows the proportion of actual negatives that the model correctly identified. So, a high specificity means that the model is good at avoiding false-positive errors.

stakeholder. A stakeholder refers to anyone who has an interest in or is affected by a particular decision or action. This could be a person, a group, or an organization. In the context of AI in education, stakeholders could include learners, parents, teachers, school administrators, policymakers, AI developers, and more.

supply-chain. A term used in cybersecurity to denote all the components that make up an IT product or service, including systems, networks, and software. Anything that could affect the

functioning of the product or service is considered part of the supply chain.

technical debt. Technical debt refers to the concept in software development where shortcuts or temporary solutions are taken during the development process to meet deadlines or deliver features quickly. These shortcuts can accumulate over time and result in suboptimal code quality, system performance, or security vulnerabilities. Just like financial debt, technical debt must be repaid eventually through refactoring, optimizing, or rewriting code to improve the system's overall quality and maintainability.

threat modeling. Threat modeling involves identifying, understanding, and prioritizing potential system threats, commonly used in cybersecurity.

tokens. In AI language processing, tokens are units of text, such as words or subword segments, used to analyze meaning. Tokenization choices affect model performance.

transparency. In the context of AI in education, transparency refers to clarity and openness about how AI systems are built, how they work, and how they make decisions.

urbanicity. Urbanicity refers to the extent and characteristics of a region's urban development, often described in terms of population density and the level of infrastructure and services available. It is used to classify geographic areas into categories such as urban, suburban, or rural.

References

Aigner, D. J., del Ángel, M., & Wiles, J. (2024). Statistical approaches for assessing disparate impact in fair housing cases. *Statistics and Public Policy*, *11*(1), Article 2263038. https://doi.org/10.1080/2330443X.2023.2263038

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA. https://www.testingstandards.net/open-access-files.html

Arras, L., Horn, F., Montavon, G., Müller, K.-R., & Samek, W. (2016). Explaining predictions of non-linear classifiers in NLP. In P. Blunsom, K. Cho, S. Cohen, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, & S. W. Yih (Eds.), *Proceedings of the 1st workshop on representation learning for NLP* (pp. 1–7). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-1601

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of*

Artificial Intelligence in Education, *32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. http://fairmlbook.org

Buijsman, S. (2023). Navigating fairness measures and trade-offs. *AI and Ethics*, *4*, 1323–1334. https://doi.org/10.1007/s43681-023-00318-0

Carey, A. N., & Wu, X. (2023). The statistical fairness field guide: Perspectives from social and formal sciences. *AI and Ethics*, *3*, 1–23. https://doi.org/10.1007/s43681-022-00183-3

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*. https://doi.org/10.1038/s41598-022-07939-1

Clemmensen, L. H., & Kjærsgaard, R. D. (2023). *Data representativity for machine learning and AI systems*. arXiv. https://arxiv.org/abs/2203.04706

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In S. Matwin, S. Yu, & F. Farooq (Eds.), *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806). ACM. https://doi.org/10.1145/3097983.3098095

Cyber & Infrastructure Security Agency. (2023a). *K-12 Education Technology Secure by Design Pledge*. https://www.cisa.gov/securebydesign/k-12-education-technology-pledge.

Cyber & Infrastructure Security Agency. (2023b). *Secure by design*. https://www.cisa.gov/securebydesign

Cyber & Infrastructure Security Agency. (2023c). *Shifting the balance of cybersecurity risk: Principles and approaches for secure by design software*. https://www.cisa.gov/sites/default/files/2023-10/SecureByDesign_1025_508c.pdf

Ding, S., & Koehn, P. (2021). *Evaluating saliency methods for neural language models*. arXiv. https://arxiv.org/abs/2104.05824

Elkan, C. (2001). The foundations of cost-sensitive learning. In B. Nebel (Ed.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence, IJCAI 2001* (973–978). Morgan Kaufmann.

High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-

trustworthy-ai

High-Level Expert Group on Artificial Intelligence. (2020). *Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment*. European Commission. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (129–145). Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates.

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, *59*(3), 338–361. https://doi.org/10.1111/jedm.12335

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (142–164). Routledge.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kruskal, W., & Mosteller, F. (1979a). Representative sampling, I: Non-scientific literature. *International Statistical Review*, *47*(1), 13–24. http://www.jstor.org/stable/1403202

Kruskal, W., & Mosteller, F. (1979b). Representative sampling, II: Scientific literature, excluding statistics. *International Statistical Review*, *47*(2), 111–127. http://www.jstor.org/stable/1402564

Kruskal, W., & Mosteller, F. (1979c). Representative sampling, III: The current statistical literature. *International Statistical Review*, *47*(3), 245–265. http://www.jstor.org/stable/1402647

Kruskal, W., & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895–1939. *International Statistical Review 48*(2), 169–195. http://www.jstor.org/stable/1403151

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30*. Curran Associates. https://papers.nips.cc/paper_files/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R., & Wendler, C. (2021). *Best practices for constructed-response scoring*. ETS. https://www.ets.org/pdfs/about/cr_best_practices.pdf

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments* (Research Report No. RR-92-39). ETS. https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1992.tb01470.x

Miao, W., & Gastwirth, J. L. (2013). Properties of statistical tests appropriate for the analysis of data in disparate impact cases. *Law, Probability and Risk*, *12*(1), 37–61. https://doi.org/10.1093/lpr/mgs032

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence centered design* (Research Report No. RR-03-16). ETS. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

OECD. (2019a). *Recommendation of the council on artificial intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD. (2019b). *Scoping the OECD AI principles: Deliberations of the expert group on artificial intelligence at the OECD (AIGO)*. https://doi.org/10.1787/d62f618a-en

Office of Educational Technology. (2023). *Artificial intelligence and future of teaching and learning: Insights and recommendations*. U.S. Department of Education. https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf

Pan, W., Cui, S., Bian, J., Zhang, C., & Wang, F. (2021). *Explaining algorithmic fairness through fairness-aware causal path decomposition*. arXiv. https://arxiv.org/abs/2108.05335

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., &

Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence* (NISTIR 8312). U.S. Department of Commerce, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

Russell, S., Perset, K., & Grobelnik, M. (2023). *Updates to the OECD's definition of an AI system explained*. OECD. https://oecd.ai/en/wonk/ai-system-definition-update.

Schwartz, R., Vassilev, A., Greene, K. K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (Special Publication 1279). U.S. Department of Commerce, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

Suk, Y., & Han, K. T. (2023). A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning. *Journal of Educational and Behavioral Statistics*, *49*(2), 151–172. https://doi.org/10.3102/10769986231171711

UNESCO. (2022). *Recommendations on the ethics of artificial intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. (2023). *Guidance for generative AI in education and research*. https://doi.org/10.54675/EWZM9535

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press. http://www.jstor.org/stable/j.ctvjf9vz4

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T.K.S., & Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, *15*, 3454–3469. https://doi.org/10.1109/TIFS.2020.2988575

Yao, L., Haberman, S. J., & Zhang, M. (2019). Penalized best linear prediction of true test scores. *Psychometrika*, *84*(1), 186–211. https://doi.org/10.1007/s11336-018-9636-7

Zapata-Rivera, D., & Arslan, B. (2024). Learner modeling interpretability and explainability in intelligent adaptive systems. In F. Santoianni, G. Giannini, & A. Ciasullo, A. (Eds.), *Mind, body, and digital brains*: *Vol. 20. Integrated Science* (pp. 95–109). Springer. https://doi.org/10.1007/978-3-031-58363-6_7

Zhu, D., Shao, X., Zhou, Q., Min, X., Zhai, G., & Yang, X. (2023). A novel lightweight audio-visual saliency model for videos. *ACM Transactions on Multimedia Computations, Communications, and Applications*, *19*(4). https://doi.org/10.1145/3576857

Suggested Citation:

Johnson, M. S. (2025). *Responsible AI for measurement and learning: Principles and practices* (Research Report No. RR-25-03). ETS. https://www.ets.org/Media/Research/pdf/RR-25-03.pdf

Action Editor: Daniel F. McCaffrey

Reviewers: Ikkyu Choi, Andrew McEachin, and Diego Zapata-Rivera

Find other ETS-published reports by searching the ETS ReSEARCHER database.