Research Article

Michael Striewe*, Sven Strickroth and Meike Ullrich

# Research challenges and future perspectives for e-assessment technologies in higher education

https://doi.org/10.1515/icom-2026-0008
Received February 23, 2026; accepted February 23, 2026;
published online March 12, 2026

**Abstract:** E-assessment technologies have rapidly evolved in higher education, transforming the evaluation of learning outcomes and the delivery of feedback to students and educators. This paper outlines the technological progression of assessment methods, from early computer-assisted systems to modern adaptive approaches powered by artificial intelligence. The main contributions of this paper are an analysis of technology drivers, current capabilities and challenges of e-assessment, as well as a structured long-term roadmap for future research directions. Based on a systematic analysis of the assessment process, key developments are identified, including advances in automatic item generation, flexible learner interaction formats, scalable feedback techniques, and personalized, adaptive assessment. At the same time, challenges remain in balancing adaptivity with data privacy, supporting diverse and authentic artifacts, and designing feedback that is both pedagogically meaningful and technically feasible. Socio-technical aspects such as trust and cultural factors add further complexity to system design. Recent progress in generative AI offers new opportunities for automation – especially in item creation and adaptive feedback – but also raises concerns regarding reliability and explainability. Finally, the article provides a forward-looking perspective on future directions and potential developments in the examined subfields of e-assessment over the next 10, 25, and 50 years.

***Corresponding author: Michael Striewe**, Department of Computer Science, Trier University of Applied Sciences, Trier, Germany,
E-mail: M.Striewe@inf.hochschule-trier.de.
https://orcid.org/0000-0001-8866-6971
**Sven Strickroth**, Ludwig-Maximilians-Universität München, Institut für Informatik, Munich, Germany, E-mail: sven.strickroth@ifi.lmu.de.
https://orcid.org/0000-0002-9647-300X
**Meike Ullrich**, Karlsruhe Institute of Technology, Institute of Applied Informatics and Formal Description Methods, Karlsruhe, Germany,
E-mail: meike.ullrich@kit.edu. https://orcid.org/0000-0003-3747-4229

**Keywords:** competence-oriented assessment; adaptive assessment; personalized assessment; automatic item generation

## 1 Introduction

In higher education, understood as formal post-secondary education at universities and comparable institutions,[1] the term “assessment” refers to a broad spectrum of methods used to evaluate, document, and provide feedback on learners’ competencies, knowledge, and skills. Assessments can serve different purposes, often termed as diagnostic assessments (to identify prerequisites), formative assessments (to monitor progress), and summative assessments (to evaluate final performance). Assessments may serve additional purposes and the terminology is not undisputed,[2] but the underlying process of conducting assessments as well as the (electronic) tools used to enact that process stay the same.

As illustrated in Figure 1, we propose that the assessment process typically comprises three abstract steps: (1) design/generation of assessment items (i.e., exercises or tasks) aligned with learning objectives, (2) learners’ work resp. interaction with the items resulting in observable artifacts, and (3) analysis of these artifacts to generate feedback (may also be a grade). The actual enactment of these abstract steps may vary, and the artifacts resulting from them may take different forms – from oral responses and practical demonstration of physical abilities to written exams.[3] All three steps may incorporate data from an existing learner profile while the latter two may also generate new insights about the learner that can be used for continuous item improvement, adaptive formats and personalized assessments.

Early e-assessment literature distinguishes between *computer-based assessment* (CBA), where both the capturing and evaluation of assessment artifacts are fully computer-supported, and the broader term *computer-assisted assessment* (CAA), where either one or both of these steps

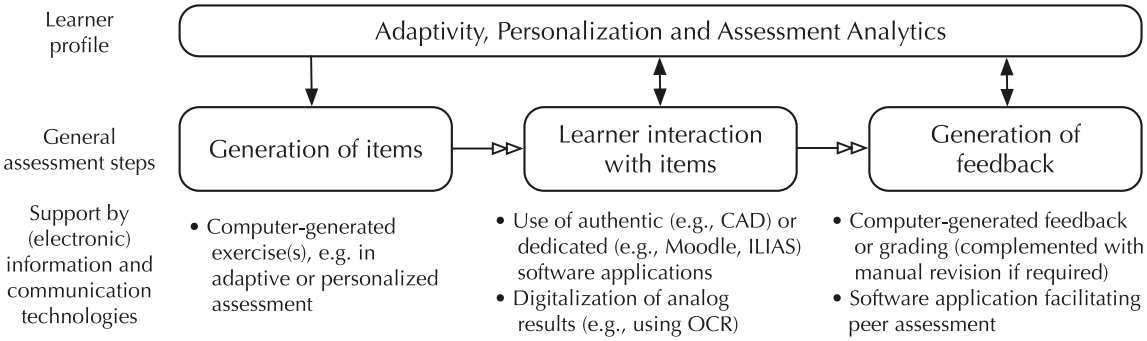


**Figure 1:** Abstract overview of the proposed assessment process used throughout this paper to illustrate the main assessment steps and possible e-assessment applications.

involve computer support.[4] Here, we use an even broader understanding: any use of (electronic) information and communication technology to support one or more of the outlined three assessment steps qualifies as e-assessment. Manual activities may still be involved, such as creation of items or feedback by a educator using a web-based quiz in a learning management system, resulting in hybrid approaches. Figure 1 lists examples of digital support for each assessment step. Overall, e-assessment is a broad field that encompasses a wide range of technologies, methods, and application contexts. Not every aspect of assessment must be digitalized – digital technologies should be applied only where they actually add value, whether in terms of achieving didactic objectives or by improving the efficiency and manageability of the assessment process. For example, advances in the assessment process may support the evaluation of a broad range of competencies and subject areas that have proven to be difficult to assess using traditional methods.

This article aims to develop a well-informed and courageous vision on the near and far future of e-assessment in higher education. It does so by providing an overview of the history and state of the art of e-assessment technologies (see Section 2) and sketching a structured long-term roadmap with next steps and open research questions related to the previously defined assessment steps (see Section 3).

In contrast to general discussions about the use of generative AI in teaching (e.g., refs. 5 or 6 on Computer Science Education), we specifically focus on research questions regarding assessment systems, without aiming to predict general developments in higher education. Our emphasis lies on the technical and procedural aspects of the assessment process itself, distinguishing our work from broader pedagogical, curricular, or cultural perspectives on assessment (e.g., refs. 7–9).

## 2 History towards the state of the art

Already 100 years ago in 1926, Pressey [10, p. 374] anticipated the potential of so-called teaching machines to relieve educators from repetitive duties like delivering instruction and providing feedback, in order to "lift from her […] shoulders as much of this burden […] and make her free for those inspirational and thought-stimulating activities which are, presumably, the real function of the teacher." With the introduction of computers in higher education in the early 1960s, actual e-assessment technologies began to emerge, including systems for programming assignments,[11] quizzes,[12] and automated essay scoring.[13] In its early phases, educational technology was primarily employed to enhance the efficiency of instructional processes.[14,15] As new technologies emerged, they increasingly influenced the evolution of learning methods, prompting the development and refinement of pedagogical approaches that both responded to and leveraged the affordances of these innovations.[14,15] Hence, technological advancement can be seen as a constant transformative force in education and enabler of (new) forms of interaction and learning.

Two main drivers have consistently fueled developments in e-assessment since then: the availability of new technologies like the Internet (e.g., ref. 16) or large language models (e.g., ref. 17) and the practical demands of educational institutions, particularly in response to massification and resource constraints (e.g., refs. 18, 19). While technological progress enables novel forms of assessment, it does not automatically align with educational goals. Thus, a third driver has become increasingly relevant: insights from educational theory and pedagogy that shape assessment processes. One prominent example is the

pedagogical principle of Constructive Alignment, which stresses coherence between intended learning outcomes, instructional activities, and assessment formats.[7] Recent work demonstrates how this principle can be explicitly operationalized in digital learning and assessment systems by systematically linking learning outcomes, teaching and learning activities, and assessment tasks.[20] Despite its widespread influence in educational theory, its application in digital assessment practice has remained limited to date. In many cases, there is a notable gap between learning objectives that target higher-order thinking skills and the assessment formats used to evaluate them. Qualitative syntheses further indicate that this gap often manifests in the continued reliance on standardized or easily automated assessment formats that emphasize recall and surface learning, even when curricula explicitly aim at analysis, application, or transfer of knowledge.[21] This mismatch underscores the need for more competence-oriented, authentic, and feedback-rich assessment systems. To fully realize the potential of e-assessment, it is essential to understand this historical trajectory.

In the following, we distinguish three major lines of development in the evolution of e-assessment technologies together with notable examples:

- **Formative e-assessment systems focused on learner interaction and feedback.** Early developments in this area aimed to support learning through immediate feedback and data-informed interventions.[22,23] Technologies such as Audience Response Systems, personalized web-based tutorials (e.g., CAPA[24]), and Intelligent Tutoring Systems (ITS)[25] exemplify this trend. These systems often emerged from computer science and AI research, with impressive technical capabilities but limited pedagogical integration. Although some ITS have shown learning gains comparable to human tutors (e.g., refs. 26–28), their design frequently lacks alignment with curriculum-level learning objectives. This limitation has been identified as a recurring issue in e-assessment research, particularly where digital assessment technologies are introduced without systematic curricular and pedagogical alignment.[21]
- **Summative and diagnostic e-assessment with a focus on psychometric validity.** A second line of development centers on formal assessments for certification or selection purposes. Computerized Adaptive Testing (CAT) and models based on Item Response Theory (IRT) provide powerful tools to ensure test efficiency and statistical reliability (e.g., refs. 29, 30). These approaches excel in standardization and scalability, but they are typically limited to closed item formats and provide little support for discipline-specific or open-ended assessment needs.
- **Domain-specific assessment tools and authentic tasks.** Domain-specific e-assessment systems have emerged to address subject-specific requirements and enable the evaluation of practical skills. Examples include the automated grading of student ALGOL programs,[31] semi-automated analysis of CAD files[32] or the automated assessment of Java programming tasks.[33] These approaches show promise in supporting Constructive Alignment by linking disciplinary practices with assessment formats, however, their use is mainly sporadic and lacks broader institutional integration, with most tools being narrowly tailored to highly specific domains or use cases.

Based on a systematic literature review, Deeva et al.[34] report over one hundred fully implemented automated feedback systems across various domains. Domain-specific reviews include even more systems, e.g., more than 170 e-assessment systems in programming education[35,36] and over one hundred approaches for the automated assessment of conceptual models in information systems.[37] These studies are cited to illustrate the scope of existing e-assessment systems across different domains, rather than providing an exhaustive census, as they are limited to scientifically published systems and do not account for proprietary or practice-driven solutions, which are likely to further increase these numbers. While these systems increasingly offer personalized and adaptive feedback, the field remains fragmented. Many implementations are not publicly accessible, and only few are designed with student-centered pedagogical principles in mind.[34] This reinforces the need for integrative design approaches that balance technical feasibility, psychometric quality, and didactic coherence.

Taken together, the evolution of e-assessment has been largely shaped by what is technically feasible and practically demanded. To meet the requirements of modern, competence-oriented higher education across disciplines, future developments must integrate pedagogical principles such as Constructive Alignment from the outset and move beyond isolated technical solutions. This includes embracing more diverse interaction formats, authentic tasks, and domain-specific needs. To operationalize Constructive Alignment in e-assessment means making the relationships between learning objectives, task design, and assessment criteria (more) explicit for both students and educators.

This requires interdisciplinary collaboration in which pedagogical considerations, disciplinary practices, and technical development are systematically aligned.

# 3 Next steps: vision and challenges

In the following sections, relevant subfields regarding the initially presented assessment process (see Figure 1) are identified and discussed individually. Nevertheless, these subfields are closely interrelated. For instance, design choices in automatic item generation (see Subsection 3.1) directly affect learner interaction with assessment artifacts (see Subsection 3.2), while both constrain how automated feedback can be generated (see Subsection 3.3). The last two subfields, adaptivity, personalization and assessment analytics (see Subsection 3.4) and socio-technical aspects (see Subsection 3.5), address cross-cutting concerns.

## 3.1 (Automatic) item generation

In simple cases, the preparation of an assessment requires to create a small set of assessment items that are carefully aligned with the intended learning outcomes. While the effort required for a single assessment may be manageable, the cumulative effort can be significant due to the frequent occurrence of assessments in higher education. In more complex cases like adaptive, personalized, or large-scale assessments, large item pools with appropriate selection mechanisms are required. If items in an item pool tend to get outdated or otherwise unusable very quickly, on-demand creation of assessment items for each individual assessment session can serve as an alternative. In any of these cases, automated item generation can be helpful, but its focus is largely different depending on the kind of assessment: Summative and diagnostic assessments require item generation processes that produce clean and unambiguous items with precise control over their psychometrical properties (i.e. their difficulty), while formative assessments require processes that produce an appealing variety of items that may (or even must) differ in difficulty and that allow for valuable feedback that promotes learning. These different goals constitute some research challenges, since advances in one area cannot easily be transferred into the other.

### 3.1.1 Generation of psychometrically valid items

Currently, there are two established approaches for automated item generation.[38,39] One approach is based on so-called item models: First, a cognitive model structure is created that covers problems, scenarios, facts, and data sources of the respective domain. From these, relevant elements and constraints can be extracted to produce an item model in the second step that defines text templates and placeholders. Actual items can be generated from this item model in the third step by filling in variables while honoring the constraints. The other approach is the cognitive design approach, which also quantifies the level and the source of cognitive complexity in an item and thus makes construct validity more explicit. While these approaches differ in complexity and control over the psychometric properties of the resulting items, both approaches try to make sure that generated items are valid by construction and thus only measure what they are supposed to measure. That helps to reduce the cost of item validation and makes automated assessments more efficient. However, the cost of item model creation is still an issue.[40] Therefore, these approaches are challenged today by automated item generation with generative AI,[41] which is also able to create psychometrically valid assessment items.[42] It creates new challenges in turn, since reliability and explainability may decrease,[17] resulting in more effort to validate and categorize items for the use with item banks, as well as a decreased applicability in fully automated settings.

Another consequence may arise for item banks and the primary goal of item generation: As of today, large item banks of validated items are a big asset for institutions involved in large-scale psychometric assessments. Research on item banking dates back to the 1950s and computerized item banking is available since the 1970s.[43] If large amounts of items with specific psychometric properties can be automatically created on demand, the need for long-living and well-maintained item banks may decrease. However, the repeated use of items from item banks does not only improve their calibration, but also creates opportunities for long-term studies and comparability of assessments across time and space. Comparability is an important aspect for high-stakes exams that must follow national standards (like medical exams) and cannot be easily guaranteed if new items are generated for every assessment. This raises the question of whether automatic item generation should focus more on personalization for formative assessments by using individual learner data throughout an on-demand generation process, since the need for generating new, static items for summative assessments drops rapidly if large item banks are available.

Within 10 years, we should be able to answer the question of how to combine the strengths of generative AI with the strengths of established item generation models to generate high-quality items at acceptable costs. That will also

allow us to see and understand the changes in the value of item banks. Still, we assume that personalization of the item generation process will become an important aspect and cannot be solved in short-term. We assume that it will take up to 25 years of research to integrate knowledge on learners into the automatic generation process in a psychometrically valid way. As of today, simple tools for automatic item generation frequently ignore psychometric properties for practical reasons. Ideally, research will be able to advance the applicability and ease of use of automatic item generation for psychometrically valid items in such a way, that in 50 years each and every item that is generated automatically also comes with a clear description of its psychometric properties.

#### 3.1.2 Generation of more diverse forms of items

Psychometric assessments follow a numerical and reductionist approach that focuses on reliability and construct validity, while modern assessment follows a broader approach and also values properties like fairness, defensibility and credibility.[44] While modern assessment uses a large variety of item forms, established item generation approaches focus on (but are not limited to) the generation of multiple-choice items and are in general typically not usable for open items.[45] In contrast, current AI-based approaches are not limited to multiple-choice items, and are able to generate assessment items of virtually any type (e.g., ref. 46), including the generation of illustrative images or alike. That makes them more interesting for formative assessments, in which it may e.g., be more important to generate items that are appealing to individual learners to increase their motivation. However, general AI-based approaches may be insufficient to capture the specifics of a particular domain, at least if it is applied in more specialized subjects in higher education. Extensive modeling of domain knowledge may still be required in the same way as it is common for item models for psychometrical assessment.

Another important challenge is the increasing diversity in learner interaction with e-assessment systems (see Subsection 3.2). New artifact types and input devices may also require new forms of item generation, larger models and more powerful algorithms to work on them. From the pure technical perspective, there may be little difference between e.g., generating an illustrative figure for a classical assessment and generating a detailed model for 3d-printing a physical assessment object. Nevertheless, the process of item generation and validation may get more complex due to physical properties as a new class of parameters in item models.

Within 10 years, we expect to see general progress in the design of (semi-)automated item generation processes that ease the creation of high-quality items with diverse item types. On that way, there are two research directions: (1) How to create proper item models or similar structures that are generally applicable to an item type or even a class of item types, and (2) how to improve the domain specific generation of high-quality items with item types that are specific for that domain? The latter will naturally touch the aspect of domain-specific interactions and artifacts. We assume that it will take up to 25 years of research until we are able to properly integrate the generation of physical objects for assessments into the general automatic item generation process, while ensuring the validity of items. In alignment with our vision for the generation of psychometrically valid items, we assume that in 50 years there will be no major difference in the automatic generation of items in diverse forms and with our without physical objects.

#### 3.1.3 Generation for automated grading and feedback

One of the primary advantages of closed items in general and multiple-choice items in particular is the ease of grading. Essentially the whole point of psychometrical assessment with Item Response Theory and alike is the fact that valid results can be obtained simply by counting correct and wrong responses due to the careful construction of the items.[47] Hence, automatic item generation in that application area by design avoids several of the general challenges of feedback generation as we will discuss in Subsection 3.3. However, the situation becomes more complex when considering other item types or elaborate and adaptive feedback. In those cases, the actual generation of an item must be accompanied by preparations for automated grading and feedback generation. That may involve the generation of a sample solution and rationales,[48] or the generation of feedback rules and texts specifically tailored to the item. Details on feedback generation are discussed in Subsection 3.3.

Within 10 years, we expect significant progress in the automatic preparation of grading and feedback generation for several item types due to current advances in generative AI. While the non-deterministic nature of generative AI still poses challenges to its use for direct grading, it seems to be fairly easy to derive deterministic grading rules and tailored feedback texts from (annotated) sample solutions with the help of generative AI. However, using sample solutions may not always be feasible, particularly for open-ended item types. Thus, we assume that it will take up to 25 years to improve automated item generation for open-ended items prepared for automated generation of elaborative feedback.

In 50 years, the preparation of grading and elaborate feedback should be an natural part of any automatic item generation process. In the same way, as psychometric properties will be part of the output from any generation process, rationales, sample solutions, feedback texts and grading rubrics will be as well.

## 3.2 Interaction with e-assessment systems

Today, a (single) learner typically interacts with e-assessment items using keyboard and mouse at a computer, or pointing and stroking at a smartphone or a tablet. Hence, a (persistent) assessment artifact is created digitally. Assessment artifacts may also be created in an analog form (using simple devices such as pen and paper) and are then digitized (e.g., using OCR or tools that detect multiple-choice marks) for further processing. These two extremes show that learners interact either directly or indirectly with an e-assessment system. Furthermore, in both cases only the final product is assessed. However, there are also training systems, such as the robot dance trainer,[49] in which only the performance is evaluated and no final product is created, or an orthopedic surgery trainer[50] that automatically assess trainees' force and movements. Hence, the assessor needs to observe the whole performance. Capturing the process and combining it with the product (if any) for assessment are both challenging.

This section systematizes assessment artifacts and their integration into the assessment system, examines implications for user interface design, and concludes with a discussion of collaborative assessments and educator interfaces.

### 3.2.1 Assessment artifacts and their way into the assessment system

Assessment artifacts span a wide range, from a simple response to a multiple choice question, a text, an image/sketch, or an authentic complex artifact created using specialized software such as CAD software. In general, artifacts must be understood quite broadly to cover all possible types of possible (and authentic) assessments. They can be classified as either products or processes/performances in principle. This distinction is important, because the creation of the final product may result from multiple retries whereas the process considers the learner's complete behavior such as which steps were performed, in what order, and how. For example, in a surgery, it is not sufficient for a trainee to simply achieve the correct anatomical result; the process of maintaining consistent instrument control and avoiding contamination of the surgical field are equally critical. These aspects of performance cannot be assessed if only the final anatomical outcome is evaluated.

In general, knowledge of the process/performance, and not only the final product, may not only be important for summative assessment, but also for formative assessments to provide targeted feedback. Today, processes are typically recorded by storing intermediate results or tracing user activities in a system.

Furthermore, assessment artifacts can be either physical or digital/virtual. A significant question is, how the artifact is created, modified, or digitized. Here, two fundamental different approaches can be differentiated: Input devices that are actively used by learners for artifact creation and manipulation, and sensors that passively capture interactions or performances or that digitize analog products.

Analog artifacts need to be digitized for processing in e-assessment systems. Therefore, sensors capturing the process and/or the final product are required. On the one hand, there are generic sensors such as video cameras, webcams, image scanners, and microphones that do not restrict learners in the tools they use for completing their assignment, in principle. For example, video-based assessments are quite common in medical education.[51] To support their analysis, there are approaches that automatically analyze body movements (e.g., ref. 52). Smart watches have also been used to assess wrist tremor during neurosurgical simulations.[53] Another example is an automatic coaching system for climbing enthusiasts based on data gathered using a smart watch as a sensor[54] which may be used by physical education teachers. On the other hand, specialized sensors may be used to digitize a product (e.g., to scan a carved sculpture using a LiDAR sensor), to capture body movements using worn sensors (e.g., inertial sensors in ref. 55 or electromyography in ref. 56), or to capture the handling of typical tools of the domain (e.g., badminton racket equipped with gyroscopes in ref. 56). Overall, sensors allow to assess authentic situations and complex (physical) workflows, but they may be perceived as too invasive and specialized sensors may not be available or affordable for everyone. Furthermore, sensors may also improve the accessibility for disabled people. Nevertheless, designing proper sensors and capturing the behavior of learners such as the handling of (domain specific) tools remains challenging.

Most digital products are created using generic haptic input devices such as mouses, keyboards, and touch interfaces on smartphones/tablets. Note, some input devices may complicate certain activities such as sketching diagrams or working with mathematical formulas compared to pen and paper, but can also simplify activities such as reorganizing a text. There are also more specialized input devices such as a

digital pen or graphic tablets that are special for certain use cases and/or help to overcome limitations of generic input devices. However, these may not be available or affordable for learners. Finally, there is an overlap of specialized haptic input devices and sensors when specific tools are required to conduct an assessment. Examples range from using 3d haptic devices (with force feedback, e.g., in ref. 50 for controlling medical instruments), a flight simulator, a digital keyboard to assess a piano play performance, to special built simulators for surgery training (e.g., ref. 57) or a paint spray gun as for a VR-based simulator for vocational education in vehicle painting.[58] Here, these devices are technically input devices to capture data, but conceptually they can be better categorized as special sensors. Such approaches seem to be rather rare today as building such customized input devices resp. sensors often poses its own challenges. We are not aware of many examples in higher education (most originate from physical education), but the examples should demonstrate their applicability, for example, in medical, teacher training, or engineering education. Another special case is the use of Augmented Reality (i.e., a camera as a generic sensor for gesture detection), when learners need to create/modify a virtual artifact. Overall, further research on suitable input devices (for authentic interaction) is needed.

Combinations are possible in more complex tasks where an analog artifact is first captured/digitized and then the digital/virtual representation is modified/analyzed by learners – in certain situations also the other direction may be possible, e.g., by 3d-printing a digital artifact. Given the absence of known examples, this approach warrants further investigation. Even when working with a digital artifact, it may be necessary to collect sensor data to cover processes outside the computer, despite the aforementioned challenges. Lastly, it is not always feasible or desirable to work with physical artifacts as they may be too expensive (e.g., only available at a specific location), unavailable in sufficient numbers, too hazardous to work with, or unsuitable for recreating specific training scenarios (e.g., medical, flight, or chemistry education). In such cases (VR) simulations and digital artifacts provide viable alternatives (e.g., refs. 57, 59). The challenge is, however, to make the interaction as authentic as possible and being able to assess all relevant aspects. This means that in addition to the essential assessment artifact(s), the e-assessment system may also collect further (sensor) data such as heart beat rate for adaptivity features (see Subsection 3.4) and/or for cheating detection/prevention.

When creating digital artifacts, there is a spectrum from accepting unstructured data such as images or audio recordings to well-defined machine-readable formats. The latter can be easily assessed automatically (see Subsection 3.3) but the former often pose significant challenges to extract and interpret the relevant data. For audio recordings, free AI-based software for transcriptions is already widely available. When dealing with images or videos, there seem to be two options: First, multimodal AI models are making progress in “transcribing”/interpreting images, and there also is research on logical scene description based on optical sensors for more than a decade now.[60] In some cases, there are ways to transform unstructured data into well-defined machine-readable formats (e.g., FeeDi interprets images of UML diagrams[61]). Second, special purpose algorithms may be developed or AI may be trained to directly work on the unstructured data. Nevertheless, well-defined machine-readable formats typically require specialized software or a special user-interface that can be embedded into generic e-assessment systems and there are no established standards (despite IMS LTI for managing the transition into different systems), yet.

Within 10 years, we envision that more special input devices are used and more authentic assessments using sensors will be possible (due to the use of AI). Furthermore, we expect significant advances in accessibility in e-assessment systems for disabled people. Within 25 years, we expect e-assessment systems to support domains with significant “offline” workflows that are recorded using generic and specialized sensors. Within 50 years, we envision that no special devices are necessary any more: Learners are observed by cameras in their authentic environment, the video is automatically evaluated and this way the actions are assessed. Nevertheless, additional sensors may be required for enhanced adaptivity and ethical questions need to be considered (see Subsection 3.5).

### 3.2.2 Design of the learners’ user interface

While specialized software or editors in e-assessment platforms can be used to solve authentic assignments, they also have an impact on the possible solution space. For example, professional UML editors may prevent specific errors (such as using a wrong line type) or limit creativity, because they only allow syntactic correct diagrams to be created. Also, having to use a specialized editor may increase the cognitive load needed to solve an assignment (compared to pen-and-paper).[62] Additionally, when existing (proprietary and non-extensible) software is used, no feedback on the process may be possible within the tool. Nevertheless, a specialized editor can also structure the solution space by providing a workflow or asking for certain intermediate results. This may help learners solve an assignment, but it may also force them to follow a specific procedure. Possible solutions to this tension may include more intuitive and adaptive/intelligent

interfaces (see Subsection 3.4). Overall, the design of the assessment interface has an impact on learner behavior: First this concerns learners' expectations what is expected as a solution (e.g., the size of text boxes has an impact on the length of the text entered[63]). Second, this may concern learners' performance when they are restricted in their options, e.g., they cannot take notes or make sketches easily. Allowing additional analog artifacts may be a solution that need to be integrated into the assessment then. Therefore, designing the interface is a challenging task, and user experience and users' actual behavior must be considered from early on.

Apart from interfaces to engage with assessments, Learning Analytics dashboards can be used to help learners reflect on their progress.[64]

Within 10 years, we expect more research on usability improvements to free cognitive capacity for the actual assessed task and on the design of actual helpful dashboards. Within 25 years, we expect more specialized software to be integrated into assessment workflows whereas the assessment and feedback also considers the process within that software (e.g., using screen capturing or standardized APIs).

#### 3.2.3 Collaborative group interactions

Most e-assessment systems nowadays seem to focus on a one-on-one relationship with a learner – assessment of collaborative group interactions are not known to us besides simple approaches analyzing quite well-structured Git history in programming practicals (e.g., ref. 65). This indicates a significant research gap. Note that the mentioned challenges in the previous sections are amplified in group assessments: When learners collaborate as individuals, their contribution and their processes need to be identified and individually acknowledged.

Over the next 10 years, we expect that e-assessment systems to support automatic assessments of collaborative group work will become more common and it will take at up to 25 years before less structured group work can be assessed automatically. Within 50 years, we envision that group work can be assessed the same way as single learners' performances.

#### 3.2.4 Interfaces for assessment management and grading

Not only learner interact with e-assessment systems, but there also are other roles such as educators (lecturers, tutors, teaching assistants, etc.), instructional designers, and domain experts. Relevant activities cover the entire course of assessments (see our framework depicted in Figure 1) depending on the specific usage scenario. Activities may be fully manual, automated, or supported by technology (hybrid). First, the user interface may need to support educators, instructional designers, and domain experts in (co)designing new assessment items (inclusive feedback, rubrics, and quality assurance) or configuring automated item generation to tailor it to a specific use case. An e-assessment system may support collaboration of the involved designers and educators by orchestration workflows (e.g., ref. 66), may employ advanced automated support approaches (e.g., quality checks[67]), or allow domain experts to use domain-specific languages/approaches to specify items (e.g., ref. 68). Second, educators and instructional designers may be supported in selecting relevant assessment items for a specific curriculum or course. While students are working on the items, a system may provide a dashboard to make learning progress and potential students at risk visible for manual intervention (e.g., ref. 69). Once learners submitted their solutions, the system may assess them fully automatically, may pre-correct them to support educators (human-in-the-loop approach), or may help educators to orchestrate other non-automated approaches such as peer review – details are discussed in Subsection 3.3. Finally, the system may provide detailed statistics on the items that can be used for evidence-based quality assurance of the items (see Subsection 3.4), and interfaces to campus management systems to report grades.

Few papers seem to have been published about interfaces for different roles. Within 10 years, more research should be conducted to evaluate needs of non-learner interfaces of e-assessment systems to better support and make them more usable for them.

### 3.3 (Automated) generation of feedback

Timely and personalized feedback is recognized as one of the most powerful drivers of learning,[70] playing a pivotal role in effective formative assessment. It demands feedback which adapts specifically to learners' needs, but many "automated feedback technologies can be considered teacher-oriented rather than student-oriented, in the sense that the first focus lies on feedback automation" [34, p. 27]. This means, the focus is on reducing the effort requested from educators for giving feedback instead of increasing the precision of feedback with respect to learners' errors. Such systems thus act in clear contrast to existing knowledge, since pedagogical research has shown that automated learning systems with error-specific try-again feedback can significantly improve learning performance.[71,72] A recent

review even suggests that interdisciplinary research with neuroscience is required "to create developmentally appropriate and individually adaptive learning environments" [73, p. 1].

#### 3.3.1 Detection of correct answers

The gap between beneficial effects of elaborated feedback and weaknesses in existing assessment systems is based on a large amount of open research questions. In open-ended assessment items, challenges start by finding out whether an answer is correct, because a simple comparison to a set of sample solutions is not sufficient. Classical approaches to automated feedback generation for open-ended items use constraints or rules that capture properties of learners' input and compare those to expected properties, such as the concept of "minimal meaningful units" for graph-based diagrams.[74] The definition, deterministic detection and comparison of such properties may be highly domain-specific, which makes it hard to generalize results. Sometimes, statistical approaches and classification via machine-learning are used due to shortcomings of rule-based approaches.[75] However, there is a class of items known as "ill-defined problems" in which the solution correctness cannot be verified automatically.[76]

These challenges are amplified by the increasing variety of learner interaction with assessment systems (see Subsection 3.2), which results in a large variety of input artifacts requiring analysis. Every educator knows the difficulties in reading handwriting and these difficulties still exist if automated assessment systems do accept handwritten text[77] or videos of a learner's performance. As a consequence, an automated system must make sure that *the system's understanding of the artifact is correct*, before it can start to judge whether the *artifact is correct*. At the same time, the increasing variety of learner interaction and input artifacts must be understood as a key enabler for elaborate feedback: Some type of feedback (such as process-related feedback in contrast to product-related feedback) is only possible if not only the final artifact of an activity is captured and assessed. Instead, the whole process of creating the artifact must be recorded and analyzed to come to the root cause of errors. Notably, not all approaches try to judge the correctness of an answer directly as in fully automatic assessment systems. Instead, semi-automatic systems can help educators to cluster large amounts of solutions for manual feedback generation and provide recommendations for appropriate feedback (see ref. 78).

Within 10 years, we should be able to answer the question on how to detect correct answers in most forms of static assessment artifacts, i.e. not only in machine-readable artifacts, but also in pictures of handwritten texts or hand drawings, photos or video frames. Within 25 years, we should also be able to answer the question on how to detect correct answers in any kind of dynamic assessment artifacts.

#### 3.3.2 Detection of specific mistakes in wrong answers

Even for closed assessment items (where it is easy to find out whether a given answer is correct), it is hard to determine the actual error automatically, and the same applies to open-ended items. Classical approaches can be used to define and detect common errors, flaws and misconceptions, so that specific feedback can be associated with each of them.[79] More sophisticated approaches take additional steps to come from observed (missing) features of an answer to root causes in form of missing competencies or misconceptions (e.g., ref. 80) for which more fundamental feedback can be generated. In any case, these approaches are specific to a domain or even a single item and thus do not scale well.

Due to the highly domain-specific nature of this aspect, we can only expect small, local advances for single domains within the next 10 years. In particular, it remains unclear if and how current advances in AI can be of help here, since the challenge is less in the task of detecting complex patterns, as in understanding the pedagogical relations between observed patterns and actual mistakes. We assume that it will be possible to formalize such relations in structured domains (like math or computer programming, where first approaches exist[81,82]), so that the detection of specific mistakes in wrong answers can be fully automated in these domains within the next 25 years.

#### 3.3.3 Generation of actual feedback

Once correct answers and mistakes are identified, actual feedback can be generated in terms of texts, pictures, and grades presented to the learner. The required form of feedback often depends on the type and context of the assessment as well as on the didactic concept of the related course or class. There is a large body of research on the different forms and pedagogical implications of feedback (see refs. 70, 83–85). In an ideal world, it is primarily a pedagogical choice what type of feedback to use within an e-assessment system, but current systems are often very limited in the feedback options they offer (e.g., common feedback types used in systems for programming education[86]).

At the same time, e-assessment systems are by their nature somewhat limited in giving feedback if compared to human capabilities. The ideal of a Socratic dialogue, in

which an educator is primarily a mentor who asks questions that stimulate reflection and consequently enlightenment in learners, is hardly reached in any operational or experimental e-assessment system today. Closest are intelligent tutoring systems that incorporate learner models and pedagogical models and are thus able to make informed choices about the feedback they present based on the learner's individual learning history. Notably, the quality of feedback then partially depends on the correctness and completeness of the learner model, that may or may not include information about e.g., the individual learning objectives, previous knowledge, or the individual learning materials seen by a particular student. However, the ideal situation may not be reached by human feedback as well, since a recent survey revealed that a significant share of studies on automatic feedback showed "no evidence that manual feedback is more efficient than automatic feedback" [87, p. 1]. Nevertheless, some of the ideas for a closer interaction between learners and systems outlined in Subsection 3.2 above can be exploited here: for example, if a system captures processes in addition to products, it can generate different feedback if some mistake occurs for the first time or repeatedly. However, determining the best reaction is still primarily a topic for psychological research and thus requires interdisciplinary work. Plain technical advances in e-assessment systems can only provide better tools to support whatever turns out to be efficient.

Within 10 years, it should be possible to make well-researched feedback types like try-again-feedback with a limited number of tries standard options in formative e-assessment systems, giving assessment authors a freedom of choice in using results from psychology in their assessment design. At the same time, it should be possible to explore rich forms of feedback beyond textual messages, including the generation of pictures, specific videos and alike, which respond directly to the inputs made by learners. This will eventually lead to more conversational-style feedback mechanisms that are well known from chat-bots and conversational agents, without having their implications explored for e-assessment systems, yet. A proper integration will require a lot of psychological studies and thus creates enough open research challenges for the next 25 or even 50 years.

### 3.3.4 Timing of automated feedback

Even if error-specific feedback can be generated automatically for every input and can be personalized for every learner, that still does not result in perfect assessment systems: A remaining challenge is the timing of feedback. Many current systems are (implicitly or explicitly) submission-oriented, which means that specific actions from learners such as clicking a button start the feedback generation process. However, timing of feedback can use three strategies from a conceptual point of view: immediate feedback, delayed feedback and on-demand feedback.[88,89] Immediate feedback is provided automatically as soon as a mistake or undesirable behavior is detected. It may be provided too early when learners are still trying to solve the issue themselves. Furthermore, it may be imprecise as some relevant information was not yet observed. Delayed feedback is provided after a learner finished working on (part of) an artifact. It can be more detailed due to a deep analysis of the artifact, but it may not be relevant to the learner any more as the thought process is not present any more. Nevertheless, it can also be seen as an additional repetition of learning material. On-demand feedback can be requested by learners themselves when they think they need feedback. However, research shows that feedback is not requested by learners who may benefit most.[90,91] Moreover, situations might occur in which learners request feedback, but in which it might be best to let them try out themselves without additional input. This is closely related to a general issue known as "gaming-the-system" – learners misusing a system's feedback functions to request as much feedback as possible to solve assignments with minimal own mental effort and consequently without actually learning anything.[92] Here, ways need to be investigated on how to detect and prevent such behavior. Notably, the general problem of timing feedback is not specific to e-assessment systems, but also appears in human teaching and is still unsolved.[93]

Within 10 years, we expect to see more diverse input to e-assessment systems (see Subsection 3.2) which will in turn produce more data that can be analyzed to decide on the proper timing of feedback. We should also be able to identify and use additional sources of information that may help to improve the timing of feedback, although they are not generally relevant to assessment. Consequently, we should be able to improve existing models for timing feedback or to come up with detailed rule-based or AI-based models for timing feedback within the next 25 years.

### 3.3.5 Alternatives to automated feedback generation

Despite the vast amount of research on automated grading and feedback generation, e-assessment systems do not necessarily need to automate this part of the assessment process. Instead, they can focus on automating other components or support scalable, non-automatic techniques for feedback such as self-assessment and peer review.

Self-assessment is "the act of monitoring one's processes and products in order to make adjustments that deepen learning and enhance performance" [94, p. 10]. Learners may use rubrics or checklists (e.g., to grade their own work or performance), or they may write an open-ended critique of their own work, performance or understanding.[94] Some authors also classify automatically generated feedback as self-assessment (e.g., refs. 95, 96).

Peer review is a learner-centered educational activity in which people with similar competencies (and equal status) evaluate fellow learners' work or processes.[97,98] It is a reciprocal approach in which learners provide and receive feedback. This method is often used in (essay) writing assignments in higher education,[97,99] but it is widely applicable and can be used when the tasks to be reviewed require some creativity – even in ill-defined domains (see ref. 76). Therefore, it can provide elaborate feedback to learners that is currently beyond the capabilities of automated technology.

As a general benefit of peer review, learners can learn from others' solutions, exchange ideas, and learn to critique work created by peers.[100,101] Furthermore, peer feedback can encourage active engagement, contribute to the development of professional skills, enhance writing and argumentation abilities, and can promote critical thinking, evaluation, decision-making, self-regulation, and communication skills.[97] Notably, providing feedback seems to lead to greater learning gains than merely receiving it, especially when learners offer elaborate, explanatory, and constructive comments.[102] There also is evidence that seeing other solutions and providing feedback can improve self-assessment.[103] Therefore, peer review cannot easily be replaced by just receiving good automatic feedback.

Standard learning technologies like wikis and blogs as well as specialized software for peer reviews have been used to orchestrate, scale, support, and evaluate peer reviews.[97,104] Technology can provide domain-specific hints and may also integrate gamification to increase motivation. Overall, online peer feedback tools have been found to have positive effects of student learning and motivation.[97,105] Emerging research also incorporates AI to support and augment the peer review process, e.g., to suggest revisions.[106,107]

On the technological side, we expect more diverse artifacts to be assessed (see Subsection 3.2) within the next 10 years and that different feedback modalities (such as audio/video or XR recordings) and more domain-specific applications as well as AI-based support features are used to support learners in providing good reviews and assessing the(ir) work. Within 25 years, we expect peer review systems to automatically match learners in a way they get the most benefits out of reviewing other (most dissimilar?) solutions and systems to automatically adapt to individual learners to tailor support features to their specific needs (e.g., to provide personalized rubrics/worked examples, or specific hints) for self- and peer-assessments based on psychology research results.

### 3.4 Adaptivity, personalization, and assessment analytics

Adaptivity and personalization were one of the main motivations behind the development of Intelligent Tutoring Systems simulating human tutors by monitoring learners' problem solving process step by step.[25] There is plenty of research indicating that ITS' can be as effective as human tutoring.[26–28]

In general, all interactions with e-assessment systems create data traces – interactions of learners but also activities that educators perform on such systems. Such data does not only include learners' submitted solutions, previous attempts and attached meta-data such as accessed material, points, grades, or feedback given by educators, but also information gathered across different learning systems, and user preferences. Moreover, additional data sources–such as information stored in campus management systems, academic calendars, and physiological signals (e.g., heart rate) collected via wearable sensors–may provide valuable insights into learners' affective and emotional states. These data streams hold significant potential for enhancing personalization and adaptivity in educational software. Note, that not only the data of a specific learner, but also the data of other learners may be helpful for finding clusters or patterns of e.g., common issues or communication. However, this data is often not used systematically, and if it is, it is mostly used within a single system. In general, the collection and analysis of data in educational contexts is called Learning Analytics (LA) or Educational Data Mining (EDM) – depending on whether the learner and interventions respective technical approaches are in the focus.[108]

Overall, there is a wide spectrum for possible personalization approaches (see ref. 19). They can concern the

- item selection/generation, e.g., based on detected misconceptions, knowledge gaps, emotional state, or context (e.g., no assignments that require loud speaking in a library),
- generation/selection of feedback, e.g., based on similar submissions or common/shared misconceptions,
- user-interface, e.g., based on the location, knowledge or device used to synchronize the current progress/settings to multiple devices, to enable on

screen formula transcription, to provide translations in the learner's native language, or to enable more advanced options which may distract beginners.

While adaptation and personalization of system behavior and user interaction may be possible purely by designing an appropriate system, adaptivity and personalization with respect to assessment contents requires the availability of a large question bank or the ability to generate new assessment items "on the fly" (see Subsection 3.1). Current adaptivity approaches are often data or expert-driven, while models based on "emotional states" (e.g., allowing for adaptive assessments that do not over- or underwhelm learners or apply an desired amount of pressure) are rare.[34]

Adaptivity and personalization are not limited to formative learning scenarios. They can also be used for individually tailored on-demand exams. Challenges are, however, fairness in terms of equivalence and equal treatment.[109] Current research knows many different effects of personalization on fairness: personalized learning may pose threats to test fairness,[110] while using prior information about students to vary surface characteristics of assessment items may increase assessment validity.[111]

Apart from the learners, personalization can also target educators, instructional designers and (student) teaching assistants. It can help them directly with the assessment such as providing pre-corrections or using collaborative approaches to ease feedback creation.[78] A noteworthy subfield of Learning Analytics is Assessment Analytics where primarily assessment data is used (on both the small and larger scale) to get insights that are hardly possible without technology.[112] From an institutional perspective, assessment analytics plays a central role in quality assurance by enabling evidence-based evaluation and continuous improvement of assessment items, formats, and processes. Examples are psychometric properties such as the difficulty or the discriminatory power of assessment items or interactions with assessment items to optimize them. Furthermore, data analysis may allow educators, instructional designers, researchers and developers of e-assessment systems to analyze how the learners interact with the system and/or with each other, identify usage and learning patterns that can help to adjust teaching methods, and to optimize the learning environment including the used software.

For ITS, quite early commonly used architectures (consisting of learner, domain, and tutor models) evolved.[28] Furthermore, there are established approaches for storing learner data in learner models (e.g., using overlay and perturbation models).[28] For general e-assessment systems, this has not happened yet despite there are (generic) commonly used components.[113] Evolving standards such as xAPI (Experience API, also known as Tincan API) or IMS/1EdTech Caliper standardize data formats and APIs allowing for storing and exchanging learning related data. Nevertheless, there are still challenges regarding the extensibility and vocabulary: The Caliper specification is quite strict and is defined by a consortium. In contrast, xAPI only defines a data format, making it quite generic. This allows developers to define their own vocabulary, but it also creates challenges, such as fragmentation through similar but different definitions.[114]

Within 10 years, we expect significant progress of adaptivity and personalization in (formative) assessments (in terms of assessment items and feedback). This includes more research on how to integrate and make use of sensor data for adaptivity. Furthermore, adaptivity features also significantly support tutors and educators in detecting learners-at-risk, grading and providing feedback. Within 25 years, we envision that adaptive exams get shorter, because only the "right" questions are asked and on-demand exams, where learners can decide when and where to take an exam, are established. Furthermore, e-assessment systems are able to provide adaptive and personalized tutoring in many domains. Within 50 years, no exams are necessary in many scenarios any more, because e-assessment systems take the role of mentors who accompany the learning process and, therefore, can also certify skills.

## 3.5 Socio-technical aspects

E-assessment systems operate within complex socio-technical environments, shaped by the interplay of humans and technical systems in an organizational context. Their design is further influenced by pedagogical, ethical, and cultural considerations. From a socio-technical perspective, manifold aspects are relevant beyond pure functionality and usability. These include trust and fairness, privacy and transparency, academic dishonesty, the social nature of learning, and cultural appropriateness.

Trust in the correctness, objectivity, and fairness of assessment systems is essential. This includes trust in the feedback provided by a system, particularly when it is generated automatically (see Subsection 3.3), as it may be subject to algorithmic bias.[115] Learners may perceive feedback as biased or inaccurate, especially in open-ended or complex tasks. In turn, learners may also consider it somewhat risky to ask for human feedback and prefer automated feedback.[116] Research from the EU-funded project TeSLA highlights how transparency and well-designed feedback mechanisms can help foster trust in adaptive e-assessment systems.[117] The use of explicable artificial

intelligence may also increase trust but "is not a full solution".[118] The need to prevent plagiarism and the feeling that it is "okay to make mistakes" are further critical components of a trust-enabling design. Educator-in-the-loop approaches with explicit human oversight may seem like a solution. However, research shows that people tend to accept outcomes "mechanically" without questioning them after they have been deemed good in many previous examples.[119,120] This automation bias – humans' tendency to favor suggestions from automated decision-making systems[121] – can lead to errors going unnoticed and undermine critical evaluation.

Privacy is a major concern, especially when e-assessment systems collect detailed learner data or include (invasive) monitoring technologies such as camera surveillance and extensive logging (see Sections 3.2 and 3.4) as digital/digitized artifacts can be persisted easily. This includes emerging modalities such as video-based responses or sensor-driven behavior tracking, which provoke fundamental ethical questions about surveillance and the right to informational self-determination. Although sensitive data needs to be protected, these practices may undermine trust and (perceived) autonomy, may influence learners behavior to experiment less, and therefore must be carefully evaluated in terms of necessity, proportionality, and transparency. Drachsler and Greller emphasize that trust in learning analytics and assessment systems can only be achieved if privacy and transparency are carefully balanced.[122] Frameworks such as the DELICATE checklist[122] provide guidance for responsible data practices.

Issues of academic dishonesty, including contract cheating and undisclosed tool use, are not specific to e-assessment systems, but have existed equally in traditional paper-based examinations.[123] In response to concerns about generative AI in unmonitored settings, discussions have included a return to conventional, ostensibly AI-proof assessment formats such as supervised in-class or technology-free exams. However, such approaches are increasingly viewed as a temporal (i.e., a current snapshot) and incomplete response, as they risk narrowing the range of knowledge and skills that assessments are intended to capture. Framing e-assessment primarily through such control-oriented responses therefore risks misplacing the problem.[124] From a socio-technical perspective, future e-assessment can instead be understood as shifting from attempts preventing the use of technological tools toward the design of assessment settings that acknowledge how students learn in digitally enriched environments, including the use of external resources, collaboration, and digital tools e.g., through so-called "AI-resistant assessment".[125]

Automatic e-assessments systems and advances in automatic assessment quality may tempt institutions to delegate all teaching/learning/assessment activities to electronic systems. One striking advantage of technology is that it has everlasting patience with learners and learners can also interact "anonymously" with it without having the feeling to get exposed to other people. Nevertheless, these properties and/or providing too much feedback may encourage learners to exploit system capabilities to solve assignments without actually learning anything.[92] Moreover, learning is fundamentally a social process. Therefore, the extent to which technology replaces and enhances human teaching represents a negotiation process within society. This is particularly an issue, as AI systems may be able to mimic human behavior and a social learning process. At the same time, it is also not an option to minimize the use of technology in favor of social interaction as a general principle. For example, it is an ethical principle that higher education institutions cannot effort to *not* use data from learning analytics[126] and a similar principle may make sense for the use of technology-enhanced assessment.

Cultural and contextual factors also play a crucial role. Learners' help-seeking behaviors, communication styles, and attitudes toward assessment can differ significantly across cultures. For example, Ogan et al. found that help-seeking patterns in ITS were not easily transferable between cultural contexts.[127] This calls for adaptable interfaces and culturally sensitive design approaches. In line with this, also the concept of inclusive design becomes increasingly relevant: it advocates for assessment systems usable across a wide range of physical, cognitive, cultural, and technological contexts. Lucke and Castro[128] present a structured process model that supports inclusive design as a means to systematically address socio-technical complexity.

Finally, several research questions remain, including how to foster trust in automated feedback, balance personalization and privacy and enable learners to own their data. Further questions concern supporting collaborative learning while preserving fairness and accountability, and integrating technology in ways that complement rather than replace human interaction. Addressing these challenges requires interdisciplinary approaches, inclusive design, and stakeholder dialogue. Socio-technical considerations are not peripheral, but central to the sustainable and equitable design of future e-assessment systems.

Over the next 10 years, we expect more systematic integration of socio-technical concerns into the design of e-assessment systems. Privacy-preserving personalization, transparent feedback mechanisms, and accessible user interfaces will become standard components of assessment

systems. Ethical guidelines and stakeholder involvement will be more commonly embedded into development processes. Within 25 years, we envision e-assessment systems being able to balance automation and human interaction dynamically. The way this balance is shaped will be determined by societal or human choices and priorities. Systems will better adapt to learners' cultural, emotional, and social contexts and provide meaningful support for collaborative learning. Personalized transparency mechanisms will be available, enabling users to understand and control how their data and feedback are generated and used. Boundaries between assessment and analytics may shift – but their direction will vary across cultures and contexts. Looking 50 years ahead, we expect socio-technical design to be integral within educational technologies and provide holistic learning environments that seamlessly support individual and collaborative learning. Systems will enable culturally adaptive assessment across the globe, while preserving individual agency and privacy, incorporating safeguards to prevent political or state systems from exploiting such technologies in ways that conflict with democratic principles or reduce individuals to fully transparent, controllable entities in an Orwellian sense.

# 4 Conclusions

This article has outlined the evolution of e-assessment from early computer-assisted systems to today's adaptive, competence-oriented approaches, showing how current technologies already enable automated item generation, flexible interaction formats, personalized feedback, and adaptivity. Our findings emphasize both the technical capabilities and the remaining limitations of present systems, especially regarding validity, inclusivity, and alignment with pedagogical goals. Against this backdrop, calls for large-scale, personalized learning supported by technology have shaped educational visions for decades. Among them, Pea's formulation captures the ambition particularly well:

> The endgame is personalized cyberlearning at scale for everyone on the planet for any knowledge domain. [129, p. 17]

Realizing this vision requires more than just technical infrastructure. Our analysis highlights three interrelated dimensions that shape the further development of e-assessment systems: technological innovations (e.g., in item generation, assessment recording, and adaptivity), pedagogical and psychological principles (e.g., Constructive Alignment and feedback theory), and socio-technical aspects (e.g., trust, privacy, inclusion). Each dimension brings both opportunities and challenges demanding attention from researchers, practitioners, learners, and policymakers alike.

This article focuses on the capabilities and challenges of e-assessment. However, note that not every aspect needs to be automated – hybrid approaches in which technology and humans work together are also desirable.

Ultimately, the future of e-assessment lies not only in automation and efficiency, but in its meaningful integration into diverse educational contexts – supporting both individual learning and collective trust in assessment outcomes. With continued interdisciplinary research and responsible system design, the field could move significantly closer to realizing Pea's vision within the next 25–50 years.

**Research ethics:** Not applicable.
**Informed consent:** Not applicable.
**Author contributions:** All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
**Use of Large Language Models, AI and Machine Learning Tools:** AI tools (Claude, Grammarly) were used to improve the language and grammar of the article.
**Conflict of interest:** The authors state no conflict of interest.
**Research funding:** None declared.
**Data availability:** Not applicable.

# References

1. OECD. *Education at a Glance 2025: OECD Indicators*; OECD Publishing: Paris, 2025.
2. Newton, P. E. Clarifying the Purposes of Educational Assessment. *Assess. Educ. Princ. Pol. Pract.* **2007**, *14* (2), 149—170.
3. Striewe, M. Lean and Agile Assessment Workflows. In *Agile and Lean Concepts for Teaching and Learning: Bringing Methodologies from Industry to the Classroom*; Parsons, D.; MacCallum, K., eds.; Springer Singapore: Singapore, 2018; pp. 187—204.
4. Conole, G.; Warburton, W. A Review of Computer-Assisted Assessment. *Res. Learn. Technol. (ALT-J)* **2005**, *13* (1), 17—31.
5. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; Kasneci, G. Chatgpt for Good? on Opportunities and Challenges of Large Language Models for Education. *Learn. Indiv. Differ.* **2023**, *103*, 102274.
6. Prather, J.; Leinonen, J.; Kiesler, N.; Benario, J. G.; Lau, S.; MacNeil, S.; Norouzi, N.; Opel, S.; Pettit, V.; Porter, L.; Reeves, B. N.; Savelka, J.; Smith, D. H. IV; Strickroth, S.; Zingaro, D. Beyond the Hype: A Comprehensive Review of Current Trends in Generative Ai Research, Teaching Practices, and Tools. In *2024 Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE 2024*; ACM: New York, NY, USA, 2025; pp. 300—338.

7. Biggs, J. B. Enhancing Teaching Through Constructive Alignment. *High. Educ.* **1996**, *32* (1), 347—364.
8. Ainsworth, L. *Rigorous Curriculum Design: How to Create Curricular Units of Study that Align Standards, Instruction, and Assessment*; Lead+Learn Press, 2011.
9. Kennedy, K. J.; Chan, J. K. S.; Fok, P. K.; Yu, W. M. Forms of Assessment and Their Potential for Enhancing Learning: Conceptual and Cultural Issues. *Educ. Res. Pol. Pract.* **2008**, *7* (3), 197—207.
10. Pressey, S. L. A Simple Apparatus Which Gives Tests and Scores - and Teaches. *Sch. Soc.* **1926**, *23* (586), 373—376.
11. Hollingsworth, J. Automatic Graders for Programming Classes. *Commun. ACM* **1960**, *3* (10), 528—529.
12. Bitzer, D. L.; Braunfeld, P. G. Computers, Teaching Machines, and Programmed Learning — Computer Teaching Machine Project: Plato on Illiac. *Comput. Autom.* **1962**, *XI* (2), 16, 18.
13. Page, E. B. The Use of the Computer in Analyzing Student Essays. *Int. Rev. Educ. / Internationale Zeitschrift für Erziehungswissenschaft / Revue Internationale de l'Education* **1968**, *14* (2), 210—225.
14. Niemiec, R. P.; Walberg, H. J. From Teaching Machines to Microcomputers: Some Milestones in the History of Computer-based Instruction. *J. Res. Comput. Educ.* **1989**, *21*, 263—276.
15. Betts, K.; Delaney, B.; Galoyan, T.; Lynch, W. Historical Review of Distance and Online Education from 1700s to 2021 in the United States: Instructional Design and Pivotal Pedagogy in Higher Education. *J. Online Learn. Res. Pract.* **2021**, *8*; https://doi.org/10.18278/jolrap.8.1.2.
16. Allen, R. The Web: Interactive and Multimedia Education. *Comput. Netw. ISDN Syst.* **1998**, *30* (16), 1717—1727.
17. Tan, B.; Armoush, N.; Mazzullo, E.; Bulut, O.; Gierl, M. A Review of Automatic Item Generation Techniques Leveraging Large Language Models. *Int. J. Assess. Tools Educ.* **2025**, *12* (2), 317—340.
18. Bull, J.; McKenna, C. *Blueprint for Computer-Assisted Assessment*; Routledge: London, 2004.
19. Strickroth, S.; Bry, F. The Future of Higher Education is Social and Personalized! Experience Report and Perspectives. In *Proc. 14th Int. Conf. on Computer Supported Education*; SciTePress, Vol. *1*, 2022; pp. 389—396.
20. Nguyen, P. H.; Tangworakitthaworn, P.; Gilbert, L. Towards Self-Regulated Individual Learning Path Generation Using Outcome Taxonomies and Constructive Alignment. In *Proceedings of the IEEE International Conference on Engineering, Technology and Education (TALE)*; IEEE, 2021; pp. 465—472.
21. Bull, D. A. Impact of Curriculum Misalignment and Assessment Practices on Student Learning Outcomes in Higher Education: A PRISMA-Guided Qualitative Content Synthesis. *Int. J. Interdiscip. Res. Innov.* **2025**, *13*, (3), 65—87.
22. Nwana, H. S. Intelligent Tutoring Systems: An Overview. *Artif. Intell. Rev.* **1990**, *4* (4), 251—277.
23. Shute, V. J.; Psotka, J. Intelligent Tutoring Systems: Past, Present, and Future. Tech. Rep.; Armstrong Laboratories, 1994.
24. Kashy, E.; Sherrill, B. M.; Tsai, Y.; Thaler, D.; Weinshank, D.; Engelmann, M.; Morrissey, D. J. CAPA-An Integrated Computer-Assisted Personalized Assignment System. *Am. J. Phys.* **1993**, *61*, 1124—1130.
25. Guo, L.; Wang, D.; Gu, F.; Li, Y.; Wang, Y.; Zhou, R. Evolution and Trends in Intelligent Tutoring Systems Research: A Multidisciplinary and Scientometric View. *Asia Pac. Educ. Rev.* **2021**, *22* (3), 441—461.
26. van Lehn, K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educ. Psychol.* **2011**, *46*, 197—221.
27. Kulik, J. A.; Fletcher, J. D. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Rev. Educ. Res.* **2016**, *86* (1), 42—78.
28. Alkhatlan, A.; Kalita, J. Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments. *Int. J. Comput. Appl.* **2019**, *181*, 1—20.
29. Kubinger, K. D. Adaptive Intelligence Diagnosticum (AID): An IRT-based Intelligence Test-Battery Fulfilling Many Practitioners' Requests. *Int. J. Sch. Cogn. Psychol.* **2016**, *3*, 1—4.
30. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; University of Chicago Press, 1980. Originally published in 1960.
31. Forsythe, G. E.; Wirth, N. Automatic Grading Programs. *Commun. ACM* **1965**, *8* (5), 275—278.
32. Krüger, D. B.; Wartzack, S. Web-Based Assessment of Cad Data in Undergraduate Design Education. *Eng. Syst. Des. Anal.* **2014**, *1*, V001T08A003.
33. Insa, D.; Pérez, S.; Silva, J.; Tamarit, S. Semiautomatic Generation and Assessment of Java Exercises in Engineering Education. *Comput. Appl. Eng. Educ.* **2021**, *29* (5), 1034—1050.
34. Deeva, G.; Bogdanova, D.; Serral, E.; Snoeck, M.; De Weerdt, J. A Review of Automated Feedback Systems for Learners: Classification Framework, Challenges and Opportunities. *Comput. Educ.* **2021**, *162*, 104094.
35. Strickroth, S.; Striewe, M. Building a Corpus of Task-based Grading and Feedback Systems for Learning and Teaching Programming. *Int. J. Eng. Pedagog. (iJEP)* **2022**, *12*, 26—41.
36. Keuning, H.; Jeuring, J.; Heeren, B. A Systematic Literature Review of Automated Feedback Generation for Programming Exercises. *TOCE* **2018**, *19* (1), 1—43.
37. Ullrich, M.; Forell, M.; Houy, C.; Pfeiffer, P.; Schüler, S.; Stottrop, T.; Willems, B.; Fettke, P.; Oberweis, A. Platform Architecture for the Diagram Assessment Domain. In *Proc. Workshop on Software Engineering for E-Learning Systems (SEELS)*; CEUR-WS, Vol. *2814*, 2021.
38. Gierl, M. J.; Lai, H. The Role of Item Models in Automatic Item Generation. *Int. J. Test.* **2012**, *12* (3), 273—298.
39. Embretson, S.; Yang, X. Automatic Item Generation and Cognitive Psychology. In *Psychometrics, vol. 26 of Handbook of Statistics*; Elsevier, 2006; pp. 747—768.
40. Kosh, A. E.; Simpson, M. A.; Bickel, L.; Kellogg, M.; Sanford-Moore, E. A Cost—Benefit Analysis of Automatic Item Generation. *Educ. Meas.: Issues Pract.* **2019**, *38* (1), 48—53.
41. Chan, K. W.; Ali, F.; Park, J.; Sham, K. S. B.; Tan, E. Y. T.; Chong, F. W. C.; Qian, K.; Sze, G. K. Automatic Item Generation in Various Stem Subjects Using Large Language Model Prompting. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100344.
42. Bhandari, S.; Liu, Y.; Kwak, Y.; Pardos, Z. A. Evaluating the Psychometric Properties of ChatGPT-generated Questions. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100284.
43. Newbould, C. A.; Massey, A. J. A Computerized Item Banking System (Cibs). *Br. J. Educ. Technol.* **1977**, *8* (2), 114—123.
44. Schuwirth, L. W. T.; van der Vleuten, C. P. M. A Plea for New Psychometric Models in Educational Assessment. *Med. Educ.* **2006**, *40* (4), 296—300.

45. Song, Y.; Du, J.; Zheng, Q. Automatic Item Generation for Educational Assessments: A Systematic Literature Review. *Interact. Learn. Environ.* **2025**, *33*, 1—20.
46. Matsumori, S.; Okuoka, K.; Shibata, R.; Inoue, M.; Fukuchi, Y.; Imai, M. Mask and Cloze: Automatic Open Cloze Question Generation Using a Masked Language Model. *IEEE Access* **2023**, *11*, 9835—9850.
47. Van der Linden, W. J.; van der Linden, W. *Handbook of Item Response Theory*, Vol. *1*; CRC Press: New York, 2016.
48. Gierl, M. J.; Lai, H. Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing. *Appl. Psychol. Meas.* **2018**, *42* (1), 42—57.
49. Paez Granados, D. F.; Yamamoto, B. A.; Kamide, H.; Kinugawa, J.; Kosuge, K. Dance Teaching by a Robot: Combining Cognitive and Physical Human—Robot Interaction for Supporting the Skill Learning Process. *IEEE Rob. Autom. Lett.* **2017**, *2*, 1452—1459.
50. Bugdadi, A.; Sawaya, R.; Bajunaid, K.; Olwi, D.; Winkler-Schwartz, A.; Ledwos, N.; Marwa, I.; Alsideiri, G.; Sabbagh, A. J.; Alotaibi, F. E.; Al-Zhrani, G.; Maestro, R. D. Is Virtual Reality Surgical Performance Influenced by Force Feedback Device Utilized? *J. Surg. Educ.* **2019**, *76*, 262—273.
51. McQueen, S.; McKinnon, V.; VanderBeek, L.; McCarthy, C.; Sonnadara, R. Video-Based Assessment in Surgical Education: A Scoping Review. *J. Surg. Educ.* **2019**, *76*, 1645—1654.
52. Han, J.; de With, P. H.; Merien, A.; Oei, G. Intelligent Trainee Behavior Assessment System for Medical Training Employing Video Analysis. *Pattern Recognit. Lett.* **2012**, *33*, 453—461.
53. Berger, L.; Civilla, L.; Dodier, P.; Rössler, K.; Moscato, F. Smartwatch-Based Wrist Tremor Assessment in Neurosurgical Simulator Training. *Sci. Rep.* **2025**, *15*, 24351.
54. Ladha, C.; Hammerla, N. Y.; Olivier, P.; Plötz, T. Climbax: Skill Assessment for Climbing Enthusiasts. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13*; ACM, 2013; pp. 235—244.
55. Rose, M.; Curtze, C.; O'Sullivan, J.; El-Gohary, M.; Crawford, D.; Friess, D.; Brady, J. M. Wearable Inertial Sensors Allow for Quantitative Assessment of Shoulder and Elbow Kinematics in a Cadaveric Knee Arthroscopy Model. *Arthrosc. J. Arthrosc. Relat. Surg.* **2017**, *33*, 2110—2116.
56. Lin, K.-C.; Cheng, I.-L.; Huang, Y.-C.; Wei, C.-W.; Chang, W.-L.; Huang, C.; Chen, N.-S. The Effects of the Badminton Teaching-Assisted System Using Electromyography and Gyroscope on Learners' Badminton Skills. *IEEE Trans. Learn. Technol.* **2023**, *16*, 780—789.
57. Huber, T.; Wunderling, T.; Paschold, M.; Lang, H.; Kneist, W.; Hansen, C. Highly Immersive Virtual Reality Laparoscopy Simulation: Development and Future Aspects. *Int. J. Comput. Assist. Radiol. Surg.* **2017**, *13*, 281—290.
58. Mulders, M.; Buchner, J.; Kerres, M. Virtual Reality in Vocational Training: A Study Demonstrating the Potential of a VR-Based Vehicle Painting Simulator for Skills Acquisition in Apprenticeship Training. *Technol. Knowl. Learn.* **2022**, *29*, 697—712.
59. Sallaberry, L. H.; Tori, R.; Nunes, F. L. S. Automatic Performance Assessment in Three-Dimensional Interactive Haptic Medical Simulators: A Systematic Review. *ACM Comput. Surv.* **2022**, *55*, 1—35.
60. Puls, S.; Graf, J.; Wörn, H. Design and Evaluation of Description Logics Based Recognition and Understanding of Situations and Activities for Safe Human-Robot Cooperation. *Int. J. Adv. Intell. Syst.* **2011**, *4* (3-4), 218—227.
61. Morawetz, E.; Hahm, N.; Thor, A. Automatisierte Bewertung und Feedback-Generierung für grafische Modellierungen und Diagramme mit FeeDI. In *21. Fachtagung Bildungstechnologien (DELFI)*; Gesellschaft für Informatik e.V.: Bonn, 2023; pp. 97—102.
62. Schuessler, K.; Striewe, M.; Pueschner, D.; Luetzen, A.; Goedicke, M.; Giese, M.; Walpuski, M. Developing and Evaluating an e-learning and e-assessment Tool for Organic Chemistry in Higher Education. *Front. Educ.* **2024**, *9*, 1355078.
63. Strickroth, S. *Unterstützungsmöglichkeiten für die computerbasierte Planung von Unterricht — ein graphischer, zeitbasierter Ansatz mit automatischem Feedback*. Phd thesis; Humboldt-Universität zu Berlin, 2016.
64. Paulsen, L.; Lindsay, E. Learning Analytics Dashboards are Increasingly Becoming About Learning and Not Just Analytics - a Systematic Review. *Educ. Inf. Technol.* **2024**, *29*, 14279—14308.
65. Guttmann, M.; Karaka, A.; Helic, D. Attribution of Work in Programming Teams with Git Reporter. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1, SIGCSE 2024*; ACM, 2024; pp. 436—442.
66. Thor, A.; Pengel, N.; Wollersheim, H. Digitalisierte hochschuldidaktik: Qualitätssicherung von prüfungen mit dem e-assessment-literacy-tool eas.lit. In *Bildungsräume 2017: Delfi 2017, Die 15. e-Learning Fachtagung Informatik, der Gesellschaft für Informatik e.V. (GI), 5. bis 8. September 2017, Chemnitz*; Igel, C., Ullrich, C., Wessner, M., eds.; Gesellschaft für Informatik: Bonn, Vol. *P-273 of LNI*, 2017; pp. 179—184.
67. Moore, S.; Nguyen, H. A.; Chen, T.; Stamper, J. Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods. In *Responsive and Sustainable Educational Futures. EC-TEL 2023. Lecture Notes in Computer Science*; Viberg, O., Jivet, I., Muñoz-Merino, P., Perifanou, M., Papathoma, T., Eds., Vol. *14200*; Springer: Cham, 2023.
68. Schöbel, K.; Merker, J.; Brassel, P.; Zöllner, M.; Hain, H. Pyrope — ein codebasierter ansatz für e-assessment in mint-fächern. In *23. Fachtagung Bildungstechnologien (DELFI 2025)*; Gesellschaft für Informatik e.V.: Bonn, 2025; pp. 449—452.
69. Wiedbusch, M. D.; Kite, V.; Yang, X.; Park, S.; Chi, M.; Taub, M.; Azevedo, R. A Theoretical and Evidence-based Conceptual Design of Metadash: An Intelligent Teacher Dashboard to Support Teachers' Decision Making and Students' Self-Regulated Learning. *Front. Educ.* **2021**, *6*, 570229.
70. Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77* (1), 81—112.
71. Attali, Y. Effects of Multiple-Try Feedback and Question Type During Mathematics Problem Solving on Performance in Similar Problems. *Comput. Educ.* **2015**, *86*, 260—267.
72. Eitemüller, C.; Trauten, F.; Striewe, M.; Walpuski, M. Digitalization of Multistep Chemistry Exercises with Automated Formative Feedback. *J. Sci. Educ. Technol.* **2023**, *32*, 453—467.
73. Adiguzel, O. C.; Potvin, P.; Esen, E. The Impact of Neuroscience and Artificial Intelligence on Feedback: A Systematic Review. *Educ. Technol. Res. Dev.* **2026**; https://doi.org/10.1007/s11423-026-10589-z.
74. Thomas, P.; Smith, N.; Waugh, K. Automatically Assessing Graph-based Diagrams. *Learn. Media Technol.* **2008**, *33* (3), 249—267.

75. Wang, H.-C.; Chang, C.-Y.; Li, T.-Y. Assessing Creative Problem-Solving with Automated Text Grading. *Comput. Educ.* **2008**, *51* (4), 1450—1466.
76. Le, N.-T.; Loll, F.; Pinkwart, N. Operationalizing the Continuum Between Well-Defined and Ill-Defined Problems for Educational Technology. *IEEE Trans. Learn. Technol.* **2013**, *6* (3), 258—270.
77. Kulkarni, M.; Adhav, G.; Wadile, K.; Deshmukh, V.; Chavan, R. Digital Handwritten Answer Sheet Evaluation System. *Int. J. Comput. Appl.* **2024**, *186* (16), 9—13.
78. Strickroth, S.; Holzinger, F. Supporting the Semi-Automatic Feedback Provisioning on Programming Assignments. In *Proc. 12. Int. Conf on Methodologies and Intelligent Systems for Technology Enhanced Learning*; Springer, 2022; pp. 13—19.
79. Gutierrez, F.; Atkinson, J. Adaptive Feedback Selection for Intelligent Tutoring Systems. *Expert Syst. Appl.* **2011**, *38* (5), 6146—6152.
80. Hubwieser, P.; Krugel, J.; Talbot, M.; Striewe, M.; Goedicke, M.; Olbricht, C. Empirical Definition of Object-oriented Programming Competencies. Tech. Rep.; Leibniz Universität Hannover, 2022.
81. Goguadze, G. *Active Math - Generation and Reuse of Interactive Exercises Using Domain Reasoners and Automated Tutorial Strategies*. PhD thesis; Universität des Saarlandes, 2011.
82. Denny, P.; Luxton-Reilly, A.; Tempero, E. D.; Hendrickx, J. Understanding the Syntax Barrier for Novices. In *Proceedings of the 16th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, ITiCSE 2011, Darmstadt, Germany, June 27-29*, 2011; pp. 208—212.
83. der Kleij, F. M. V.; Feskens, R. C. W.; Eggen, T. J. H. M. Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Rev. Educ. Res.* **2015**, *85* (4), 475—511.
84. Boomgaarden, A.; Loibl, K.; Leuders, T. Fostering Learning from Errors — Computer-Based Adaptivity at the Transition Between Problem Solving and Explicit Instruction. *Journal für Mathematik-Didaktik* **2024**, *45*, 9.
85. Panadero, E.; Lipnevich, A. A. A Review of Feedback Models and Typologies: Towards an Integrative Model of Feedback Elements. *Educ. Res. Rev.* **2022**, *35*, 100416.
86. Jeuring, J.; Keuning, H.; Marwan, S.; Bouvier, D.; Izu, C.; Kiesler, N.; Lehtinen, T.; Lohr, D.; Peterson, A.; Sarsa, S. Towards Giving Timely Formative Feedback and Hints to Novice Programmers. In *Proc. Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE-WGR '22*; ACM, 2022; pp. 95—115.
87. Cavalcanti, A. P.; Barbosa, A.; Carvalho, R.; Freitas, F.; Tsai, Y.-S.; Gašević, D.; Mello, R. F. Automatic Feedback in Online Learning Environments: A Systematic Literature Review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100027.
88. Kulik, J. A.; Kulik, C.-L. C. Timing of Feedback and Verbal Learning. *Rev. Educ. Res.* **1988**, *58* (1), 79—97.
89. Corbett, A. T.; Anderson, J. R. Locus of Feedback Control in Computer-based Tutoring: Impact on Learning Rate, Achievement and Attitudes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '01*; Association for Computing Machinery, 2001; pp. 245—252.
90. Aleven, V.; Koedinger, K. R. Limitations of Student Control: Do Students Know when They Need Help? In *Intelligent Tutoring Systems. ITS 2000. Lecture Notes in Computer Science*; Gauthier, G., Frasson, C., VanLehn, K., Eds., Vol. *1839*; Springer: Berlin, Heidelberg, 2000.
91. Aleven, V.; Stahl, E.; Schworm, S.; Fischer, F.; Wallace, R. Help Seeking and Help Design in Interactive Learning Environments. *Rev. Educ. Res.* **2003**, *73* (3), 277—320.
92. Baker, R. S.; Corbett, A. T.; Koedinger, K. R. Detecting Student Misuse of Intelligent Tutoring Systems. In *Intelligent Tutoring Systems*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; pp 531—540.
93. Lohr, D.; Kiesler, N.; Keuning, H.; Jeuring, J. "Let Them Try to Figure it out First" - Reasons Why Experts (Do Not) Provide Feedback to Novice Programmers. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2024*; ACM, 2024.
94. Andrade, H. L. A Critical Review of Research on Student Self-Assessment. *Front. Educ.* **2019**, *4*, 87.
95. Baruque, B.; Herrero, Á. Self-Assessment Web Tool for Java Programming. In *International Joint Conference. CISIS 2015. Advances in Intelligent Systems and Computing*; Herrero, Á., Baruque, B., Sedano, J., Quintián, H., Corchado, E., Eds., Vol. *369*; Springer: Cham, 2015.
96. Kay, J.; Li, L.; Fekete, A. Learner Reflection in Student Self-Assessment. In *Proceedings of the Ninth Australasian Conference on Computing Education - Volume 66, ACE '07, (AUS)*; Australian Computer Society, Inc., 2007; pp. 89—95.
97. Kerman, N. T.; Banihashem, S. K.; Karami, M.; Er, E.; van Ginkel, S.; Noroozi, O. Online Peer Feedback in Higher Education: A Synthesis of the Literature. *Educ. Inf. Technol.* **2023**, *29*, 763—813.
98. Topping, K. Peer Assessment Between Students in Colleges and Universities. *Rev. Educ. Res.* **1998**, *68*, 249—276.
99. Cho, K.; Schunn, C. D. Scaffolded Writing and Rewriting in the Discipline: A Web-based Reciprocal Peer Review System. *Comput. Educ.* **2007**, *48* (3), 409—426.
100. Nicol, D. From Monologue to Dialogue: Improving Written Feedback Processes in Mass Higher Education. *Assess Eval. High Educ.* **2010**, *35* (5), 501—517.
101. Indriasari, T. D.; Luxton-Reilly, A.; Denny, P. A Review of Peer Code Review in Higher Education. *ACM Trans. Comput. Educ.* **2020**, *20* (3), 1—25.
102. Wu, Y.; Schunn, C. D. Passive, Active, and Constructive Engagement with Peer Feedback: A Revised Model of Learning from Peer Feedback. *Contemp. Educ. Psychol.* **2023**, *73*, 102160.
103. Strickroth, S. Does Peer Code Review Change My Mind on My Submission? In *Proc. Conf. on Innovation and Technology in Computer Science Education (ITiCSE 2023)*; ACM, 2023; pp. 498—504.
104. Gao, X.; Noroozi, O.; Gulikers, J.; Biemans, H. J.; Banihashem, S. K. A Systematic Review of the Key Components of Online Peer Feedback Practices in Higher Education. *Educ. Res. Rev.* **2024**, *42*, 100588.
105. Dawson, P.; Henderson, M.; Ryan, T.; Mahoney, P.; Boud, D.; Phillips, M.; Molloy, E. Technology and Feedback Design. In *Learning, Design and Technology: An International Compendium of Theory, Research, Practice and Policy*; Spector, J. M., Lockee, B. B., Childress, M. D., Eds.; Springer: Cham, Switzerland, 2018.
106. Guo, K.; Pan, M.; Li, Y.; Lai, C. Effects of an ai-supported Approach to Peer Feedback on University Efl Students' Feedback Quality and Writing Ability. *Internet High Educ.* **2024**, *63*, 100962.

107. Gao, X.; Schunn, C. D.; Noroozi, O.; Gulikers, J.; Banihashem, S. K.; Biemans, H. Consistency in Providing Peer Feedback: The Role of Individual and Situational Factors. *Assess Eval. High Educ.* **2025**, *50*, 1—15.
108. Baker, R.; Siemens, G. Learning Analytics and Educational Data Mining. *Camb. Handb. Lean. Sci.* **2014**, 253—272.
109. Denny, P.; Manoharan, S.; Speidel, U.; Russello, G.; Chang, A. On the Fairness of Multiple-Variant Multiple-Choice Examinations. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education, SIGCSE '19*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 462—468.
110. Katz, D.; Huggins-Manley, A. C.; Leite, W. Personalized Online Learning, Test Fairness, and Educational Measurement: Considering Differential Content Exposure Prior to a High Stakes End of Course Exam. *Appl. Meas. Educ.* **2022**, *35* (1), 1—16.
111. Mislevy, R. J.; Haertel, G.; Cheng, B. H.; Ructtinger, L.; DeBarger, A.; Murray, E.; Rose, D.; Gravel, J.; Colker, A. M.; Rutstein, D.; Vendlinski, T. A "Conditional" Sense of Fairness in Assessment. *Educ. Res. Eval.* **2013**, *19* (2-3), 121—140.
112. Ellis, C. Broadening the Scope and Increasing the Usefulness of Learning Analytics: The Case for Assessment Analytics. *Br. J. Educ. Technol.* **2013**, *44* (4), 662—664.
113. Striewe, M. Components and Design Alternatives in E-Assessment Systems. In *Software Architecture - 13th European Conference, ECSA 2019, Paris, France, September 9-13, 2019, Proceedings*, 2019; pp. 220—228.
114. Ehlenz, M.; Heinemann, B.; Leonhardt, T.; Röpke, R.; Lukarov, V.; Schroeder, U. Eine forschungspraktische perspektive auf xapi-registries. In *DELFI 2020 — Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V.*; Gesellschaft für Informatik e.V.: Bonn, 2020; pp. 331—336.
115. Baker, R. S.; Hawn, A. Algorithmic Bias in Education. *Int. J. Artif. Intell. Educ.* **2021**, *32* (4), 1052—1092.
116. Henderson, M.; Bearman, M.; Chung, J.; Fawns, T.; Shum, S. B.; Matthews, K. E.; de Mello Heredia, J. Comparing Generative AI and Teacher Feedback: Student Perceptions of Usefulness and Trustworthiness. *Assess Eval. High Educ.* **2025**, 1—16; https://doi.org/10.1080/02602938.2025.2502582.
117. Baneres, D.; Rodríguez, M. E.; Guerrero-Roldán, A. E. *Engineering Data-Driven Adaptive Trust-Based E-Assessment Systems*; Springer: Cham, 2020.
118. Farrow, R. The Possibilities and Limits of Explicable Artificial Intelligence (XAI) in Education: A Socio-Technical Perspective. *Learn. Media Technol.* **2023**, *48* (2), 266—279.
119. Green, B. The Flaws of Policies Requiring Human Oversight of Government Algorithms. *SSRN Electron. J.* **2022**, *45*, 105681.
120. Agudo, U.; Liberal, K. G.; Arrese, M.; Matute, H. The Impact of Ai Errors in a Human-in-the-Loop Process. *Cogn. Res.: Princ. Implications* **2024**, *9*, 1.
121. Cummings, M. Automation Bias in Intelligent Time Critical Decision Support Systems. In *AIAA 1st Intelligent Systems Technical Conference*, 2004.
122. Drachsler, H.; Greller, W. Privacy and Analytics: It's a Delicate Issue a Checklist for Trusted Learning Analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16, LAK '16*; ACM Press, 2016; pp. 89—98.
123. Newton, P. M. How Common is Commercial Contract Cheating in Higher Education? *Front. Educ.* **2018**, *3*, 67.
124. Lodge, J. M. The Evolving Risk to Academic Integrity Posed by Generative Artificial Intelligence: Options for Immediate Action. Tech. Rep.; Tertiary Education Quality and Standards Agency (TEQSA), 2024.
125. Alkouk, W. A.; Khlaif, Z. N. Ai-resistant Assessments in Higher Education: Practical Insights from Faculty Training Workshops. *Front. Educ.* **2024**, *9*, 1499495.
126. Slade, S.; Prinsloo, P. Learning Analytics: Ethical Issues and Dilemmas. *Am. Behav. Sci.* **2013**, *57* (10), 1510—1529.
127. Ogan, A.; Walker, E.; Baker, R.; Rodrigo, M. M. T.; Soriano, J. C.; Castro, M. J. Towards Understanding How to Assess Help-Seeking Behavior Across Cultures. *Int. J. Artif. Intell. Educ.* **2015**, *25* (2), 229—248.
128. Lucke, U.; Castro, T. The Process of Inclusive Design. In *Proceedings of the 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, 2016; pp. 18—22.
129. Pea, R.; Jacks, D. *The Learning Analytics Workgroup: A Report on Building the Field of Learning Analytics for Personalized Learning at Scale*; Stanford University: Stanford, CA, 2014.