Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Artificial intelligence-based assessment in education

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Fang, Ying	Central China Normal University	0000-0002-4086-5856
Roscoe, Rod D.	Arizona State University	0000-0001-8327-4012
McNamara, Danielle S.	Arizona State University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 4/20/2023

Check type of content being submitted and complete one of the following in the box below:

- ☐ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☑ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Handbook of Artificial Intelligence in Education, 485-504, Edward Elgar Publishing, Inc., Massachusetts, USA

DOI or URL to published work (if available) http://dx.doi.org/10.4337/9781800375413

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** Institute of Education Scie through **[Grant number]** R305A1900 to **Institution]** Arizona State U. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Arizona State University or the U.S. Department of Education.

21. Artificial intelligence-based assessment in education

Ying Fang, Rod D. Roscoe and Danielle S. McNamara

Artificial intelligence (AI) refers to the ability of machines to adapt to new situations, deal with emerging situations, solve problems, answer questions, devise plans, and perform other functions that require some level of intelligence typically evident in human beings (Coppin, 2004). AI methods have been increasingly used in a variety of settings such as employment, healthcare, policing, and education to assess human behaviors and influence decision making. For example, AI tools are used in today's recruitment to screen résumés, conduct interviews and analyze candidates' personality traits and skills (Upadhyay & Khandelwal, 2018; Van Esch et al., 2019). AI techniques are also widely used in healthcare to collect patients' information, process and analyze the information, help decision making, and improve resource utilization (Jiang et al., 2017; Yu et al., 2018). In the field of policing, facial recognition technology is used by police officers to identify suspects from the image data captured by cameras (Ariel, 2019; Robertson et al., 2016). Additionally, AI methods enable predictive policing, which purports to anticipate where crimes may occur, who might commit the crimes, and potential victims based on historical crime records (Ariel, 2019; Perrot, 2017). AI methods and techniques are also applied in manufacturing to facilitate planning and control, such as detecting defective products, optimizing manufacturing supply chains, and designing products (Hayhoe et al., 2019; Li et al., 2017, Singh & Gu, 2012).

With respect to education, AI techniques undergird diverse tools used to evaluate students and in turn guide learning and instruction (Corbett et al.,1997; Mousavinasab et al., 2021). One class of widely used educational technology is intelligent tutoring systems (ITSs), which provide immediate feedback and customized instruction to learners (Graesser et al., 2012; Psotka et al., 1988; Shute & Psotka, 1996; VanLehn, 2011). For such systems to be "intelligent," AI tools and techniques are used to communicate with students, collect data, conduct analyses, and make instructional decisions. In digital learning games and simulations, AI methods have also been used to provide personalized and engaging experiences to learners. AI methods are implemented in these learning environments to assess students dynamically, adjust task difficulty and learning paths, and provide cognitive feedback and motivational support (Conati et al., 2013; Hooshyar et al., 2021; Peirce et al., 2008; see Chapter 20 by McLaren and Nguyen and Chapter 9 by Aleven et al.). In the field of educational assessment, such as automated writing evaluation (AWE), AI technology is used to grade essays and provide qualitative and quantitative feedback during or after essay drafts (Burstein et al., 2013; Foltz et al., 2013; Warschauer & Ware, 2006). The common features across these educational technologies are the automated and personalized feedback, instruction, or experiences provided to students that are enabled by assessments using AI techniques.

There are many approaches for AI-based assessments in educational technologies, and some technologies use multiple approaches as means of assessments. For example, machine learning and natural language processing are used together by some technologies to analyze

language data and evaluate students' performance. Meanwhile, different approaches (e.g., collecting diverse data sources or applying different computational methods) may be used to assess the same constructs (e.g., performance or knowledge). To develop effective educational technologies, developers and educators need to make decisions about how AI-based assessments will be implemented to meet the educational goals. In turn, these decisions introduce dependencies and constraints on other assessment components. For instance, the type of data that are collected may be constrained by the student traits targeted for evaluation. Additionally, the data type may restrict the computational methods. To describe and facilitate some of these considerations, this chapter will provide a framework for AI-based assessment (AIBA) in educational technology and design.

The framework comprises five interrelated dimensions that broadly address the "purposes" and "procedures" of AIBA for educational technologies: goals, target constructs, data sources, computational methods, and visibility. Purposes broadly refer to the overarching aims of the assessment, which include two dimensions: *goals* of AIBA and *constructs* to be assessed in support of those goals. Procedures primarily refer to the means for achieving the purposes, which involve data sources, computational methods, and visibility of assessments. *Data sources* refer to the data collected to measure the constructs in a robust, valid, and reliable way. *Computational methods* are the methods used in AIBA to process, transform, and analyze the data to make inferences. Assessment *visibility* refers to whether assessments are overt and obvious to the learner or if they are covert or unobtrusive (i.e., stealth assessment).

The remainder of this chapter will further describe each of the five dimensions within the contexts of purposes and goals while using educational systems and software as examples to illustrate each dimension. We anticipate this framework to be useful to researchers, educators, and designers who value a common language for discussing AIBAs for educational technology. We also anticipate that this framework will help researchers, educators, and designers to begin thinking about the complex decisions related to implementing AIBAs in educational technologies.

THE AIBA FRAMEWORK

As shown in Figure 21.1, the framework comprises five dimensions that are nested within two overarching themes: (1) purposes, including the *goals* of assessments and the *constructs* to be assessed, and (2) procedures for achieving purposes, including *data sources*, *computational*

Figure 21.1 AI-based assessment (AIBA) framework

methods, and *visibility*. Purposes will be described first, followed by procedures, because the decisions about goals and constructs of assessments often influence decisions concerning procedures. Nevertheless, methodological innovations can inspire new purposes by making it possible for educational technologies to implement assessments (e.g., natural language processing innovations render it possible for technologies to provide feedback on certain aspects of writing). Therefore, the relationship between purposes and procedures is often reciprocal.

Purposes of AIBA

Goals

Educational technology developers, educators, and learners use AI-based assessments to support a variety of needs (see Chapter 9 by Aleven et al., Chapter 15 by Pozdniakov et al., and Chapter 23 by Ritter and Koedinger). Perhaps the most prominent or commonplace of these goals are personalized instruction and feedback (see Chapter 9 by Aleven et al.). Educational technologies can help students develop their declarative knowledge of the subject or acquire essential procedural knowledge and skills, and technologies can facilitate learners' awareness of their performance and progress (see Chapter 18 by Casas-Ortiz et al. and Chapter 19 by Martinez-Maldonado et al.).

One goal of AIBAs is to guide *personalized instruction* or *training* offered via educational technologies (e.g., to teach physics knowledge or train reading strategies). AI-based assessments can use learners' inputs, behaviors, performance, and related metrics to deliver or recommend appropriate learning materials or practice opportunities needed by individual learners. Technologies that incorporate such personalization features have been shown to be more effective in facilitating student learning than traditional instruction or computer-based tools without such features (see Kulik & Fletcher, 2016; Ma et al., 2014; Steenbergen-Hu & Cooper, 2013; VanLehn, 2011).

Personalization is a ubiquitous and defining goal for many educational technologies (see Chapter 23 by Ritter and Koedinger). Two examples are *AutoTutor for CSAL* (Center for the Study of Adult Literacy) and *Assessment and LEarning in Knowledge Spaces* (ALEKS). AutoTutor for CSAL is a web-based ITS that delivers comprehension instruction (Graesser et al., 2016; Fang et. al., 2021). A typical lesson in the system consists of a video reviewing a comprehension strategy and practices scaffolded by conversational agents. When students practice the reading strategy within a lesson, they usually begin with medium-level materials. The system subsequently branches students into different conditions where the learning materials are of different difficulty based on the assessment of their performance. ALEKS is an ITS that provides math skill training. This system tracks the knowledge states of students (e.g., the topics students know and the topics students are ready to learn) and adaptively responds with assignments that are sensitive to these knowledge states (Craig et al., 2013).

Another important goal of AIBAs is to provide *feedback* to facilitate or inform students regarding their progress. Feedback in the context of education is essential for knowledge and skill acquisition, and is thus an important element of the instruction in educational technologies (Epstein et al., 2002; VanLehn, 2011; see Chapter 9 by Aleven et al.). AI affords rapid or real-time assessments in educational technologies, which in turn enable automated, immediate, and personalized feedback. Educational assessment is typically categorized into summative assessment, which evaluates how much students learn, and formative assessment, which

assesses how students learn. The feedback generated from these two types of assessments is referred to as summative and formative feedback, respectively.

Summative feedback provides information regarding learners' current performance or performance relative to others (e.g., scores, skill bars, rankings, and completion reports). For example, Criterion is an automated writing evaluation tool developed by Educational Testing Service (Burstein et al., 2013) that generates summative feedback immediately after a student submits an essay. The feedback includes a holistic score and diagnostic feedback about grammar, vocabulary usage, mechanics, style and organization, and development.

In contrast, formative feedback includes information that helps learners appreciate current states, desired states, and ways to improve or grow (e.g., hints, prompts, and motivating messages). Durlach and Spain (2014) summarized formative feedback generated during the instruction of intelligent tutoring systems into two types: corrective feedback and supportive feedback. Additionally, within each type, there are five levels according to the purpose of feedback, the amount of information within the feedback, and the techniques used to trigger and deliver feedback. Corrective feedback includes feedback with only summary score (level 0), minimal feedback with item accuracy information (level 1), correct answers or explanation of correct answers (level 2), error-sensitive feedback (level 3), and context-aware feedback (level 4). Supportive feedback includes no support (level 0), fixed hints on request (level 1), locally adaptive hints on request or triggered (level 2), context-aware hints on request or triggered (level 3), and context-aware hints on request or triggered together with interactive dialogue (level 4).

Many systems provide both corrective and supportive feedback, but to different levels. For example, *AutoTutor* is a conversation-based learning environment that has been used to teach a wide range of topics such as computer literacy, physics, scientific reasoning, and comprehension strategies (Graesser et al., 2020; Nye et al., 2014; see Chapter 11 by Rus et al.). *AutoTutor* provides formative feedback to scaffold student learning through conversations between students and computer agents. After students answer a question, the computer agents typically provide formative feedback including whether the student's answer is positive or negative, pumps that ask students to say more or take some action, hints that guide students toward a particular answer, and prompts that get students to use a particular word or phrase (Graesser et al., 2020).

Some systems provide both summative and formative feedback. For example, *Writing Pal* is an ITS for writing that can provide formative feedback to students during their writing practice (McNamara et al., 2012; Roscoe & McNamara, 2013). After a student submits an essay, the system provides immediate formative feedback, including a holistic rating from poor to great (6-point scale), a message addressing particular writing goals and strategy-based solutions, and prompts toward relevant lessons or practice games for just-in-time strategy instruction or practice (Roscoe & McNamara, 2013; Roscoe et al., 2014). *Betty's Brain* is a computer-based learning environment built upon the "learning-by-teaching" paradigm in which students teach a virtual agent, Betty, scientific topics (e.g., ecology) (Biswas et al., 2016; see Chapter 20 by McLaren and Nguyen). The feedback generated by the system includes summative feedback such as Betty's performance on the quizzes, and formative feedback including Betty's explanations for her answers and learning strategies provided by the mentor agent (Biswas et al., 2016; see Chapter 4 by Azevedo and Wiedbusch).

Importantly, these goals (i.e., guiding personalized instruction and providing feedback) are not mutually exclusive, and many modern systems address both. For example, *Cognitive Tutor*

is a math ITS that provides step-by-step feedback to students as they work through problems (Koedinger & Corbett, 2006; Pane et al., 2014; Ritter et al., 2007). The system also evaluates students during their problem solving and reconstructs what knowledge students have already mastered versus what they have yet to learn. The system then decides the learning path (e.g., selects the problem targeting the knowledge components that are missing or in error) for each individual student (Anderson & Gluck, 2001; Koedinger & Corbett, 2006). VanLehn (2006) considered two loops while describing ITS behaviors: an outer loop at the task level (e.g., solving a mathematics problem) and an inner loop at the step level (e.g., a solution step in a mathematics problem). The outer loop involves customizing the learning paths (e.g., problem selection), and the inner loop is usually where the detailed step-by-step feedback is provided to students. Many ITSs include both loops to achieve their educational goals. It should be noted that achieving the specific goals may require different data sources, or computational methods, which foreshadows later elements of the framework.

Constructs

Constructs refer to the variables, states, or phenomena that will be assessed or measured. As alluded to earlier, the constructs are partly specified by the AIBA goals. That is, the constructs are typically important learner characteristics that potentially influence the efficacy of instruction and training (e.g., prior knowledge or skill mastery) or require feedback (e.g., on the accuracy of a solution or appropriate use of a strategy). For example, if the goal of an educational system is to provide customized math instruction and formative feedback, measuring students' math competencies is the basis for the system to time the delivery of appropriate learning materials and just-in-time feedback. Similarly, students' affect (e.g., boredom or confusion) may affect their level of engagement and efficacy of the instruction. As such, the system may also embed assessments to measure students' affective states for further motivational intervention. Constructs that are commonly measured in AIBAs include knowledge, skills, learning strategies, and learners' cognitive and affective states.

Knowledge and *skills* are perhaps the most common constructs assessed in many educational technologies. Knowledge refers to familiarity with factual information and theoretical concepts. Skill refers to the ability to apply knowledge to specific situations. Knowledge and skills are typically evaluated to determine what instruction or training to provide, what feedback to offer, and to understand whether the instruction or training provided by the technologies is effective. For example, *Why2/AutoTutor* (Nye et al., 2014) is an ITS designed to teach physics. The system dynamically assesses students' physics knowledge to provide feedback and customize instruction. *Physics Playground* is a 2-dimensional computer game helping students learn Newtonian physics (Shute et al., 2020). The assessments embedded in the game evaluate students' physics knowledge to guide the game level selection and learning support (e.g., hints and worked examples) provided by the system. *Cognitive Tutor* and *Wayang Outpost* (Arroyo et al., 2014) are ITSs to improve mathematics skills. When students interact with these systems, their mathematics problem-solving skills are continuously assessed to guide which feedback and instructional materials to provide. *AutoThinking* is an adaptive digital game designed to promote students' skills and conceptual knowledge in computational thinking (Hooshyar et al., 2021; see Chapter 20 by McLaren and Nguyen). The player takes the role of a mouse that solves programming problems to collect cheese pieces in a maze while also escaping from two cats. Students' skills and knowledge are assessed during the game, and the system adaptively adjusts the performance of the cats based on student performance.

The game also provides different types of feedback (i.e., textual, graphical, and video) to players based on the assessment of the current state of the maze and students' skill levels.

Strategies refer to intentional procedures that students know and use to improve their performance, which are also commonly assessed constructs. Strategies include domain-specific and domain-general strategies, and adaptive and maladaptive strategies. The assessment of students' learning strategies is to provide feedback, facilitate instruction or training and help improve the effectiveness of educational technologies. For instance, *MetaTutor* is an ITS designed to foster students' self-regulated learning (Azevedo et al., 2019; see Chapter 4 by Azevedo and Wiedbusch). It evaluates domain-general learning strategies such as planning, monitoring, and note-taking when students work on biology problems, and provides feedback to help students enhance self-regulated learning. *Interactive Strategy Training for Active Reading and Thinking* (*iSTART*, Boonthum et al., 2011; McNamara, 2021; McCarthy et al., 2020b) is an ITS that provides reading strategy training. *iSTART* evaluates students' comprehension strategies such as monitoring, paraphrasing, bridging, and elaboration while they read challenging texts.

Learner states such as cognitive states (e.g., cognitive load and confusion) and affective states (e.g., emotions and boredom) are more frequently assessed during the initial development of AIBAs, often to assess students' attitudes toward the instruction, training, or feedback (i.e., user experience) (Taub et al., 2021; see Chapter 4 by Azevedo and Wiedbusch and Chapter 20 by McLaren and Nguyen). These states are the signals of success and struggles and may mediate or moderate how students approach learning and assessment tasks (Jackson & McNamara, 2013). Some ITSs further incorporate algorithms to assess learner states to guide adaptive instruction. For example, *Cognitive Tutor* implements multiple algorithms to assess students' engagement by detecting their gaming behavior (e.g., repeatedly asking for help until the system reveals the correct answer, or inputting answers quickly and systematically) and off-task behavior (Baker, 2007; Baker et al., 2008a, 2008b). Similarly, *Affective AutoTutor* is a version of *AutoTutor* that detects students' affective states such as boredom, confusion, and frustration. When negative emotions are detected, the system provides empathetic and motivational statements with the goal of reengaging students (D'Mello et al., 2009; D'Mello & Graesser, 2013).

Procedures of AIBA

Procedures refer to the means of achieving AIBA purposes and goals, which involve data sources, computational methods, and visibility of assessments. Data sources refer to the data collected to measure constructs in a robust, valid, and reliable manner. Computational methods are employed to process, transform, and analyze the data to make inferences. Assessment visibility refers to whether assessments are overt and obvious to the learner, or if they are covert (i.e., explicit) or unobtrusive (i.e., implicit or stealth assessment). In short, these design dimensions broadly describe the methods and approaches used to operationalize the assessments.

Data sources

Data sources refer to information, input, or output from students that inform assessments such as knowledge, skills, strategies, and emotions. Commonly collected data sources include performance, behavior, language, and biometric data.

Performance data are commonly collected to assess student work and outcomes (e.g., completeness, accuracy, and quality), and in turn, make inferences about students' knowledge or

skills. The specific form of performance data can vary between tasks. For example, *iSTART* embeds two types of tasks. One type requires students to generate constructed responses (e.g., writing self-explanations) in sentences or paragraphs. The other type asks students to select correct answers from the provided choices. The performance data for the two types of tasks are the quality of constructed responses and correctness of the answers, respectively (McCarthy et al., 2020b). In the domain of computer science, student submission sequences on programming tasks can be collected as a type of performance data (see Chapter 13 by Mao et al.).

Another frequently collected data source are students' *behaviors*, which contain information about student actions and interactions. Behavior data include students' actions such as their keystrokes, button clicks, mouse movements, and navigation through the system as well as how they implement the actions (e.g., time interval between actions; see Chapter 15 by Pozdniakov et al.). Behavior data have often been used to infer students' learning strategies, cognitive states, or affective states in many educational systems (Aleven et al., 2016; Paquette & Baker, 2019; Snow et al., 2014, 2015, 2016; see Chapter 4 by Azevedo and Wiedbusch).

Language is a data source containing information about what and how students communicate verbally (e.g., via spoken or written input). Language products can be revealing of performance (e.g., knowledge and skills). For example, one type of *iSTART* practice requires students to generate constructed responses (e.g., self-explanations) in sentences or paragraphs. From this practice, students' reading skills can be assessed using features of these student-generated texts; the quality of the constructed responses is a performance metric (Allen & McNamara, 2015; McCarthy et al., 2020a).

Biometric data refers to the information about students' physiology, gaze, posture, and facial expressions (e.g., electroencephalogram (EEG), galvanic skin response (GSR), and eye-tracking), which is usually collected to assess learner states (Azevedo & Gašević, 2019; Cabada et. al., 2020; Pham & Wang, 2018).

Finally, when multiple data sources are used together, they are often referred to as *multimodal data* (Blikstein & Worsley, 2016; Worsley, 2018; Worsley & Blikstein, 2018; see Chapter 10 by Lajoie and Li). It is a common practice for researchers and designers to use multimodal data sources to evaluate students and provide instruction and feedback accordingly. For example, biometric data are often used together with performance data to infer students' affective states (Azevedo & Gašević, 2019; D'Mello & Graesser, 2013; Sharma et al., 2020; Wang & Lin, 2018).

Computational methods

Computational methods refer to statistical analyses, AI, machine learning, and other methods used to process and analyze data, make inferences, and automate responses to students. These methods may depend on or be influenced by data sources because some data sources demand specific methods for data processing or transformation. For example, natural language processing (NLP) methods are required to process and analyze language data. Computational methods may also be influenced by purposes. Different methods may provide output that is more or less useful depending on the types of feedback provided to students or adaptivity adopted by the educational technologies. In this section, we discuss how three categories of computational methods of AIBAs (i.e., Bayesian methods, NLP methods, and machine-learning methods) are implemented in various systems to assess the target constructs and achieve educational goals.

Bayesian methods

Bayesian methods refer to the statistical methods that use probability to represent uncertainty; they have been adopted by a wide range of educational technologies. One frequently used Bayesian method is Bayesian network analysis, which graphically represents a set of variables and their conditional independencies, and then exploits this information to reduce the complexity of probabilistic inference (Culbertson, 2016; Pearl, 1988).

Bayesian network analysis is often used to assess the mastery of a skill or a knowledge component. In the design of several ITSs, domain knowledge is decomposed by researchers into smaller units referred to as knowledge components (Conati & Zhou, 2002; Conati et al., 2018; VanLehn et al., 2005). A complex cognitive task usually involves numerous knowledge components (Anderson, 2014). For example, *Andes* (VanLehn et al., 2005) is an ITS developed to assist physics problem solving, covering about 75% of the AP Physics B curriculum. The *Andes* developers decomposed relevant content into 308 knowledge components addressing well-known principles, such as Newton's Second Law or Ohm's Law. Solving a problem in *Andes* typically involves multiple knowledge components and requires several steps. To assess a student's mastery of a knowledge component (e.g., a physics rule), the system evaluates the probability of mastery of the prerequisite knowledge and students' performance on each step while solving a problem. Specifically, *Andes* builds a Bayesian network whose nodes and links represent how the steps in a problem solution derive from previous steps and physics rules for each problem or task solved by a student. When a problem-solving step is entered in the *Andes* interface, *Andes* retrieves the corresponding node in the Bayesian network, sets its value to "true" and computes the posterior probability of other nodes in the network given this new evidence. The posterior probabilities become the prior probabilities of the nodes in the network for the next problem that uses the previous rule (Conati, 2010). As such, *Andes* dynamically assesses students' knowledge and updates the student model.

Cognitive Tutor is an ITS developed to teach math, and it also implements Bayesian methods for real-time diagnosis of students' math knowledge and skills. Referred to as *Bayesian Knowledge Tracing* (BKT; Corbett & Anderson, 1995), it is equivalent to the two-node Bayesian network (Baker et al., 2008a). BKT algorithms compute the probability of a student mastering a rule at time T_{i+1} as a function of the probability of knowing the rule at time T_i and observations of the student's performance on steps pertaining to that rule at T_{i+1}. With BKT, *Cognitive Tutor* diagnoses students' knowledge while they interact with the system, and provides individualized feedback and instruction based on the diagnosis. In addition to assessing knowledge and skills, Bayesian network analysis has also been used to estimate learner states. For example, *Prime Climb* is an educational game designed to help students learn number factorization. Bayesian networks were used to model students' affective states (e.g., joy, distress, shame) when they play the game (Conati & Zhou, 2002; Conati, 2011). *Wayang Outpost* is an ITS designed to teach high-school mathematics, and a Bayesian network was applied to infer students' attitudes toward learning (Arroyo & Woolf, 2005; Arroyo et al., 2014).

Natural language processing (NLP) methods

NLP is a broad category of methods used for different levels of natural language processing, such as speech recognition, syntactic analysis, semantic analysis, and discourse analysis (Burstein et al., 2013; Elliot et al., 2003; D'Mello et al., 2011; Litman et al., 2006; McNamara et al., 2007, 2013). Speech recognition focuses on diagramming a continuous speech signal

into a sequence of known words. Syntactic analysis analyzes groups of words conforming to the rules of formal grammar. For example, it determines the ways words are clustered into components such as noun and verb phrases. Semantic analysis focuses on understanding the meaning and interpretation of words, signs and sentence structure. It involves diagramming a sentence to a type of meaning representation such as a logical expression. Discourse analysis focuses on the nature of the discourse relationships between sentences and how context impacts sentential interpretations. NLP methods are widely used in ITSs designed for language and literacy training (e.g., reading, comprehension, and writing), and conversation-based ITSs that require students' input to be in the form of language (Dascalu et al., 2017, 2018; McNamara et al., 2018).

In addition to ITSs, NLP methods are also widely used in AWE systems to assign scores and provide diagnostic feedback. For example, the *Intelligent Essay Assessor* (IEA) developed by Pearson Education is an AWE system that can analyze students' writing and provide automated feedback (Foltz et al., 2013). The feedback generated by IEA includes a holistic score and analytic feedback on six traits: ideas (i.e., developing a main idea with supporting ideas), organization (using organization to highlight the main idea and move to the conclusion), conventions (using conventions such as spelling, punctuation, and grammar correctly), sentence fluency (using a variety of sentence lengths and structures correctly), word choice (using a variety of specific, descriptive, and appropriate words), and voice (using a consistent and effective tone). The NLP method embedded in IEA is Latent Semantic Analysis (LSA), which uses statistical computations to extract and represent the meaning of words. Specifically, given a large corpus of text with millions of words and thousands of documents, a matrix is created that indicates the context in which each word occurs. The context of a word is the document in which it occurs, which may be the sentence, paragraph, or entire text. This is a sparse matrix because most terms occur in few documents, and it is a large matrix because there are many terms across many documents. The matrix is then reduced to discover its latent properties using singular value decomposition (SVD). This process creates a multidimensional LSA space, wherein a word is represented by a fixed-size vector of real numbers. A sentence or document is also represented by a fixed-size vector, made by summing component word vectors. Words, sentences, and documents can be compared with each other by comparing their vectors. To assess the quality of essays, IEA compares the LSA vectors representing student essays with the vectors of pre-scored essays on the same topic to assess the semantic similarity. The similarity between a target essay and a pre-scored essay is measured by the cosine between the two vectors. As such, the semantic content of two essays can be compared and a score derived based on their similarity (Foltz et al., 1999; Landauer et al., 2007, 2013; McNamara, 2011).

Another example of using NLP methods for data processing and analysis is *AutoTutor*, in which conversational agents hold conversations with students in natural language and provide feedback (Graesser et al., 2004, 2020); therefore, one key data source is language. For a typical *AutoTutor* task, there are multiple conversational turns between students and computer agents. The conversations in the system are designed according to a conversational framework referred to as expectation and misconception tailored (EMT) dialogue. Specifically, for each main question there is a list of expectations (e.g., anticipated good answers and steps in a procedure) and a list of anticipated misconceptions (e.g., bad answers, incorrect beliefs, errors, and bugs) created by domain experts. As students articulate their answers over multiple

conversational turns, their answers are compared with the expectations and misconceptions using LSA. *AutoTutor* compares the LSA vector of students' answers to the vectors of the expectations and misconceptions (Graesser et al., 2020). The assessment results indicate whether an expectation is covered or a misconception exists, and also affect the next dialog move to present to students (e.g., pumps, hints, and prompts).

Machine-learning methods

Machine learning (ML) had been defined as a "field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959). Machine-learning methods refer to the computer algorithms that improve automatically through experience and data. Machine-learning methods are frequently used in AIBAs to analyze different types of data, such as language, performance, behavioral and multimodal data. Machine-learning algorithms usually build models based on sample data, known as "training data," in order to make predictions or decisions.

Machine-learning methods have been combined with NLP in many AWE systems to grade essays and provide automated feedback. The implementation of automatic essay scoring usually consists of a training stage and a scoring stage. During the training stage, NLP tools are used to identify and extract linguistic features that can predict the scores of the sample essays (i.e., training data) rated by human experts. Weighted statistical models are then trained to predict the score using the features. The linguistic features identified for the scoring are usually related to the rubrics which define how the essays should be scored. For example, *Criterion* is an AWE designed to help students develop their writing skills by providing automated and constructive feedback. *Criterion* uses the e-rater scoring engine to identify linguistic features and score essays (Burstein, 2003; Burstein et al., 2013). The linguistic features e-rater extracts consist of three modules: syntax, discourse, and topic. The features in the syntax module (e.g., subjunctive auxiliary verbs, subordinate clauses) capture syntactic variety in an essay, and they are identified by a parser. The discourse module features capture discourse-based relationships and organization in essays, and are identified using a conceptual framework of conjunctive relations including cue words (e.g., using words like "perhaps" or "possibly" to express a belief), terms (e.g., using conjuncts such as "in summary" and "in conclusion" for summarizing), and syntactic structures (e.g., using complement clauses to identify the beginning of a new argument). The topic module features capture the vocabulary usage and topical content. A vector–space model is used to convert the training essays into weight vectors which populate the training space.

Similar to *Criterion*, *My Access!* is an AWE system developed by Vantage Learning, and it uses the IntelliMetric scoring system to evaluate over 400 syntactic, discourse, and semantic features. Those features are described in five dimensions: focus and coherence, organization, development and elaboration, sentence structure, and mechanics and conventions. The features in focus and coherence dimension capture a single point of view, cohesiveness and consistency of purpose, and main ideas in an essay. The features on organization focus on an essay's transitional fluency and logic of discourse, such as the introduction and conclusion, logical structure, logical transitions, and the sequence of ideas. The features in development and elaboration dimension analyze the breadth of the content and the supporting ideas in an essay. The sentence structure category features examine sentence complexity and variety such as syntactic variety, sentence complexity, usage, readability, and subject-verb agreement. The features about mechanics and conventions describe whether the essay includes the

conventions of standard English such as grammar, spelling, capitalization, sentence completeness, and punctuation (Elliott et al., 2003; Schultz, 2013).

Although the systems differ in the approach of extracting linguistic features, the goal of training the statistical models is to accurately predict the expert-rated scores with the selected features. Next, the statistical models including selected linguistic features and their weights are fitted into the new data (i.e., essays) to assign scores and provide diagnostic feedback on a set of dimensions. For instance, e-rater provides a holistic score and diagnostic feedback about grammar, vocabulary usage, mechanics, style and organization, and development (Burstein, 2013). Similar to e-rater, IntelliMetric generates a score reflecting overall performance as well as diagnostic feedback on five rhetorical and analytical dimensions such as conventions and organization (Elliott et al., 2003; Schultz, 2013).

Some ITSs designed to help students improve their literacy skills (e.g., reading and writing) also use machine-learning and NLP methods to estimate the performance in the form of language. For example, *iSTART* implements NLP and machine-learning methods to assess self-explanations generated by students and provide formative feedback. To evaluate a self-explanation, *iSTART* compares its LSA vector with the vector of four benchmarks separately. The four benchmarks are (1) the words in the title of the passage, (2) the words in the target sentence, (3) prior words or sentences in the prior text that are causally related to the target sentence, and (4) the words that appear more than once in the previously collected explanations and do not appear in the other benchmarks. The final rating of the self-explanation is based on a weighted sum of the four LSA cosines between the explanation and the four benchmarks (McNamara et al., 2007). In addition, *iSTART* combines LSA and word-based algorithms using machine-learning methods (i.e., discriminant function analysis) to generate formative feedback that prompts the readers to improve their self-explanations (McNamara, 2021). *Writing Pal* also implements NLP and machine-learning methods to assess essays and guide the feedback and individualized training (McNamara et al., 2012; Roscoe & McNamara, 2013; Roscoe et al., 2014).

In addition to analyzing language, machine-learning methods are used for the analysis of other data sources. For example, *Affective AutoTutor* is a version of *AutoTutor* that can detect students' affective states and provide supportive feedback based on students' emotions (D'Mello et al., 2009; D'Mello & Graesser, 2013). In addition to recording students' inputs when they have conversations with the computer agents in the system log files, *Affective AutoTutor* also records students' facial features and body languages with cameras and the body posture measurement system (BPMS). The multimodal data, including conversational cues (i.e., dialogue features), facial expressions, and body language, are analyzed with machine-learning methods (e.g., Naive Bayes logistic regression and support vector machines) to classify students' emotions such as confusion, boredom, flow, frustration, and neutral emotion (D'Mello & Graesser, 2013). Similar to *Affective AutoTutor*, *MetaTutor* applies machine-learning algorithms such as Random Forests, Naive Bayes, Logistic Regression, and Support Vector Machines to analyze eye-tracking data (i.e., gaze data) and classify students' emotions (e.g., boredom and curiosity) during learning (Jaques et al., 2014).

Machine-learning methods are also used to analyze performance and behavioral data. For example, *LP-ITS Tutor* is an ITS teaching students linear programming. It adopts machine-learning methods to assess students' performance so that the system can provide individualized instruction (Abu Naser, 2012). Specifically, the log files recorded by the system contain rich information about students and their learning details, such as performance, actions, time

on task, and problem details. The log files are analyzed with the machine-learning algorithm (i.e., Artificial Neural Networks) to predict students' performance and decide what learning materials to provide to each individual student. *Cognitive Tutor* embeds machine-learning algorithms to detect students' off-task behavior and gaming behavior using performance and behavioral data such as actions and action times (Baker, 2007; Baker et al., 2008b). *iSnap* is a computer-based learning environment designed to teach a computer science course for non-majors (Price et al., 2017). This system implemented deep-learning algorithms to model student learning progression (e.g., temporal sequences in log files) and predict student success or failure to provide adaptive intervention (see Chapter 13 by Mao et al.).

Bayesian, NLP, and machine learning are widely used methods in educational technologies. However, other methods are also viable. For example, several ITSs use constraint-based modeling, such as *SQL-Tutor*, which teaches SQL database language (Mitrovic, 2003), *EER-Tutor*, which teaches conceptual database design (Mitrovic & Suraweera, 2016), and *J-LATTE*, a tutor for learning Java (Holland et al., 2009). The fundamental idea behind constraint-based modeling is that all correct solutions to a problem are similar in that they do not violate any domain principles. The constraint-based systems store a set of domain-specific constraints representing the characteristics of correct solutions. A constraint is usually in the form of "If is <relevant condition> is true, then <satisfaction condition> had better also be true." A solution is incorrect if it violates one or more constraints. As such, constraint-based modeling is primarily the match between students' solutions and constraints.

Visibility

Assessment visibility refers to whether the AIBAs implemented in educational technologies are overt and obvious to the learner, or if they are covert or unobtrusive. In most current educational technologies, students' task performances are explicitly evaluated: students usually know that the feedback they receive is based on performance assessment. For example, when students submit an essay in an AWE system and receive the score and analytical feedback immediately, it is evident that the essay is assessed by the system, and the feedback is based on the assessment. Similarly, when a computer agent tells a student whether an answer is correct, and gives some hints, students usually understand the feedback is based on an underlying evaluation.

By contrast, *stealth assessment* is a type of assessment that evaluates students covertly and unobtrusively. Stealth assessment refers to the evidence-based assessments woven directly and invisibly into gaming environments (Shute, 2011; Shute & Ventura, 2013). The data needed to assess students (i.e., players) are generated when students interact with the game, and can be used to infer students' skills or knowledge. In a well-designed game assessment scenario, students may not be aware of being assessed during the gameplay. Stealth assessment was initially proposed and explored because some competencies such as persistence, creativity, self-efficacy, openness, and teamwork were revealed to substantially impact student academic achievement (O'Connor & Paunonen, 2007; Poropat, 2009; Sternberg, 2006). However, those competencies were not assessed in the educational technologies. Researchers then proposed using performance-based assessments to assess those competencies by analyzing how students use knowledge and skills in the real world. One approach to assessing those competencies is via game-based learning environments, which can provide students with diverse scenarios requiring the application of differing competencies. When students play games, their performance, behavior and other types of data are collected, and analyzed to infer their competencies. As such, students are being assessed unobtrusively during the gameplay.

For example, in *Physics Playground* stealth assessments have been implemented to evaluate students' competencies including physics knowledge, persistence, and creativity (Shute & Rahimi, 2021; Ventura & Shute, 2013; Wang et al., 2015). Specifically, a student produces a dense stream of performance data during the gameplay. The data is recorded by the game system in a log file, which is analyzed using Bayesian methods (i.e., Bayesian networks) to infer students' competencies. The system then provides formative feedback and other forms of learning support to students during gameplay based on the assessment. Stealth assessments for performance-based measures and domain-general competencies (e.g., persistence) have been found to be valid across a variety of game environments (Ventura & Shute, 2013; Ventura et al., 2013).

Another example of stealth assessment is embedded in a game called *Use Your Brainz*, which is a slightly modified version of a popular commercial game *Plants vs. Zombies 2* (Shute et al., 2016). The stealth assessment also uses players' performance data, which are the log files recording in-game behaviors. The performance is analyzed with Bayesian methods to infer students' problem-solving skills. The stealth assessment measures based on performance data have also been validated against external measures.

In addition to assessing general competencies independent of domain, stealth assessments can also be used to assess domain knowledge and skills that are not explicitly evaluated by the intelligent systems. For example, during the self-explanation practice in *iSTART*, students receive immediate feedback on the quality of their self-explanations which is based on the NLP analysis of the self-explanations. The linguistic and semantic features of those explanations are not just signatures of self-explanation quality, they also provide windows into students' underlying comprehension skills and knowledge. Features of language provide information about individual differences in vocabulary, domain, and world knowledge as well as literacy skills. For example, rare words, complex syntax, and language that is cohesive are signatures of stronger reading skills (Allen et al., 2015, 2016a, 2016b; Allen & McNamara, 2015; McCarthy et al., 2020a). NLP provides a means to understand and evaluate language skills and knowledge because features of language (e.g., syntax, concreteness, meaningfulness, cohesion) provide proxies aligned with how students are processing, can process, are producing, and can produce language (McNamara, 2021). Stealth assessment of literacy skills has strong potential to enhance the adaptivity of systems in which students generate natural language input.

DISCUSSION

This chapter introduces the AI-based assessment (AIBA) framework, which categorizes the purposes and procedures of AIBA using five interrelated dimensions in educational technologies: goals, constructs, data sources, computational methods, and visibility.

The overarching purposes are described from two dimensions (see Figure 21.1), which are goals of AIBA, and constructs to be assessed in support of those goals. The broad goals of AIBA comprise the provision of summative and/or formative feedback to students, and guidance of personalized instruction or training. These goals are not mutually exclusive as many educational technologies address both. Constructs refer to the variables that are assessed by AIBA, which are heavily influenced by the goals. Specifically, constructs are typically important learner characteristics that potentially influence the efficacy of instruction, training, and

feedback. The commonly measured constructs include knowledge, skills, learning strategies, and learners' cognitive states and emotions.

The procedures of AIBAs refer to how the purposes are achieved in educational technologies, which involve three dimensions: data sources, computational methods, and visibility of assessments. Data sources refer to data collected in order to reliably and accurately measure the constructs. The commonly collected data sources in modern educational technologies include performance data, language, behavior, and biometric data (e.g., physiology, gaze, body language, and facial expressions). Multiple data sources are sometimes used together in the assessments, referred to as multimodal data. The access to rich data enables the assessment of some constructs that can be challenging to evaluate. Data collected within the AIBAs are processed and analyzed using appropriate computational methods to infer the target learner traits. The commonly used methods in current educational technologies include Bayesian methods, NLP, and machine-learning methods. Each category includes a variety of specific techniques for data processing and analysis. Different methods and techniques are often used together in educational tools to evaluate the target constructs and achieve their goals. Finally, visibility refers to whether the AIBAs implemented in educational technologies are obvious or unobtrusive to learners. The latter dimension includes stealth assessments, which are usually implemented in game-based learning environments seamlessly to evaluate students unobtrusively during their gameplay.

Regarding recent AIBA advances, learning analytics is a research area that has played an important role. Learning analytics researchers have collected data generated in various AI systems and explored diverse methods, particularly machine-learning methods, to analyze the data, which helps achieve the goals of AIBAs. For example, learning analytics research examined the data from digital games to better understand how students interacted with games and help improve game design (see Chapter 20 by McLaren and Nguyen). Learning analytics researchers also analyzed data collected from Massive Open Online Courses (MOOCs) to help improve the instructional design (Doleck et al., 2021; Er et al., 2019; Shukor & Abdullah, 2019). Importantly, although we can collect rich data from many sources, it is not the case that more data is always better. We should make evaluations based on the goals of AIBAs and the characteristics of learning environments to decide what data are necessary for valid and reliable assessments.

AIBAs undergird diverse educational technologies to guide students' learning and teachers' instruction during their use of the educational technologies. Designers and educators make multiple, intertwined decisions regarding the design of the instructional technologies. Each decision and design choice can impact others as they often introduce constraints in the assessments. System designers often consider these dimensions prior to developing an educational system. The AIBA framework is designed to facilitate and guide that process such that researchers and developers can discern a clearer picture of the AI-based technology prior to development. As the five dimensions of AIBA are interrelated and intertwined, we recommend that they are considered as a whole during the design and implementation of AI-based educational technologies.

REFERENCES

Abu Naser, S. S. (2012). Predicting learners' performance using artificial neural networks in linear programming intelligent tutoring system. *International Journal of Artificial Intelligence & Applications, 3*(2), 65–73.

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, *26*(1), 205–223.

Allen, L. K., Dascalu, M., McNamara, D. S., Crossley, S. A., & Trausan-Matu, S. (2016a). Modeling individual differences among writers using ReaderBench. In *EduLearn* (pp. 5269–5279). Barcelona, Spain: IATED.

Allen, L. K., & McNamara, D. S. (2015). You are your words: Modeling students' vocabulary knowledge with natural language processing. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)* (pp. 258–265). Madrid, Spain: International Educational Data Mining Society.

Allen, L. K., Perret, C. A., & McNamara, D. S. (2016b). Linguistic signatures of cognitive processes during writing. In J. Trueswell, A. Papafragou, D. Grodner, & D. Mirman (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society in Philadelphia, PA* (pp. 2483–2488). Austin, TX: Cognitive Science Society.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. In J. Baron, G. Lynch, N. Maziarz, P. Blikstein, A. Merceron, & G. Siemens (Eds.), *Proceedings of the 5th International Learning Analytics & Knowledge Conference* (pp. 246–254). Poughkeepsie: ACM.

Anderson, J. R. (2014). *Rules of the Mind*. Psychology Press.

Anderson, J. R., & Gluck, K. (2001). What role do cognitive architectures play in intelligent tutoring systems. *Cognition & Instruction: Twenty-Five Years of Progress*, 227–262.

Ariel, B. (2019). Technology in policing. In D. L. Weisburd & A. A. Braga (Eds.), *Innovations in Policing: Contrasting Perspectives* (2nd ed., pp. 521–516). Cambridge, England: Cambridge University Press.

Arroyo, I., & Woolf, B. (2005). Inferring learning and attitudes from a Bayesian network of log file data. In C. K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Twelfth International Conference on Artificial Intelligence in Education* (pp. 33–40). Amsterdam: IOS Press.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education, 24*(4), 387–426.

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, *96*, 207–210.

Azevedo, R., Mudrick, N. V., Taub, M., & Bradbury, A. (2019). Self-regulation in computer assisted learning systems. In J. Dunlosky & K. Rawson (Eds.), *Handbook of Cognition and Education* (pp. 587–618). Cambridge, MA: Cambridge University Press.

Baker, R. S. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1059–1068.

Baker, R. S., Corbett, A. T., & Aleven, V. (2008a). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In *International Conference on intelligent Tutoring Systems* (pp. 406–415). Heidelberg, Berlin: Springer.

Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008b). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*(3), 287–314.

Biswas, G., Segedy, J. R., & Bunchongchit, K. (2016). From design to implementation to practice a learning by teaching system: Betty's Brain. *International Journal of Artificial Intelligence in Education*, *26*(1), 350–364.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238.

Boonthum, C., McCarthy, P. M., Lamkin, T., Jackson, G. T., Magliano, J., & McNamara, D. S. (2011). Automatic natural language processing and the detection of reading skills and reading comprehension. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 234–239). Menlo Park, CA: AAAI Press.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated Essay Scoring: A cross Disciplinary Approach* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 55–67). New York: Routledge.

Cabada, R. Z., Rangel, H. R., Estrada, M. L. B., & Lopez, H. M. C. (2020). Hyperparameter optimization in CNN for learning-centered emotion recognition for intelligent tutoring systems. *Soft Computing, 24*(10), 7593–7602.

Conati, C. (2010). Bayesian student modeling. In *Advances in Intelligent Tutoring Systems* (pp. 281–299). Berlin, Heidelberg: Springer.

Conati, C. (2011). Combining cognitive appraisal and sensors for affect detection in a framework for modeling user affect. In *New Perspectives on Affect and Learning Technologies* (pp. 71–84). New York, NY: Springer.

Conati, C., Jaques, N., & Muir, M. (2013). Understanding attention to adaptive hints in educational games: An eye-tracking study. *International Journal of Artificial Intelligence in Education*, *23*(1), 136–161.

Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in education needs interpretable machine learning: Lessons from open learner modelling. *arXiv preprint*. https://arxi.org/1807.00154.

Conati, C., & Zhou, X. (2002). Modeling students' emotions from cognitive appraisal in educational games. In *International Conference on Intelligent Tutoring Systems* (pp. 944–954). Heidelberg, Berlin: Springer.

Coppin, B. (2004). *Artificial Intelligence Illuminated*. Boston, MA: Jones & Bartlett Learning.

Corbett, A. T., & Anderson, J. R. (1995) Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253–278.

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In *Handbook of Human-Computer Interaction* (pp. 849–874). North-Holland.

Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti, A. E., Sterbinsky, A., Cheney, K. R., & Okwumabua, T. (2013). The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers & Education*, *68*, 495–504.

Culbertson, M. J. (2016). Bayesian networks in educational assessment: The state of the field. *Applied Psychological Measurement*, *40*(1), 3–21.

Dascalu, M., Allen, K. A., McNamara, D. S., Trausan-Matu, S., & Crossley, S. A. (2017). Modeling comprehension processes via automated analyses of dialogism. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1884–1889). London, UK: Cognitive Science Society.

Dascalu, M., Crossley, S. A., McNamara, D. S., Dessus, P., & Trausan-Matu, S. (2018). Please Readerbench this text: A multi-dimensional textual complexity assessment framework. In S. Craig (Ed.), *Tutoring and Intelligent Tutoring Systems* (pp. 251–271). Hauppauge, NY: Nova Science Publishers.

D'Mello, S., Craig, S., Fike, K., & Graesser, A. (2009). Responding to learners' cognitive-affective states with supportive and shakeup dialogues. In *International Conference on Human-Computer Interaction* (pp. 595–604). Heidelberg, Berlin: Springer.

D'Mello, S., & Graesser, A. (2013). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(4), 1–39.

D'Mello, S. K., Dowell, N., & Graesser, A. (2011). Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied*, *17*(1), 1.

Doleck, T., Lemay, D. J., & Brinton, C. G. (2021). Evaluating the efficiency of social learning networks: Perspectives for harnessing learning analytics to improve discussions. *Computers & Education*, *164*, 104124.

Durlach, P. J., & Spain, R. D. (2014). *Framework for Instructional Technology: Methods of Implementing Adaptive Training and Education*. Fort Belvoir, VA: Army Research for the Behavioral and Social Sciences.

Elliott, S., Shermis, M. D., & Burstein, J. (2003). Overview of intelliMetric. In *Automated Essay Scoring: A Cross-Disciplinary Perspective* (pp. 67–70). Elbaum.

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., et al. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record*, *52*, 187–201.

Er, E., Gómez-Sánchez, E., Dimitriadis, Y., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., & Álvarez-Álvarez, S. (2019). Aligning learning design and learning analytics through instructor involvement: A MOOC case study. *Interactive Learning Environments*, *27*(5–6), 685–698.

Fang, Y., Lippert, C. Z., Chen, S., Frijters, J. C., Greenberg, D., & Graesser, A. C. (2021). Patterns of adults with low literacy skills interacting with an intelligent tutoring system. *International Journal of Artificial Intelligence in Education*. Advance online publication. DOI: 10.1007/s40593-021-00266-y.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2), 939–944.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In *Handbook of Automated Essay Evaluation* (pp. 68–88).

Graesser, A. C., Cai, Z., Baer, W. O., Olney, A. M., Hu, X., Reed, M., & Greenberg, D. (2016). Reading comprehension lessons in AutoTutor for the center for the study of adult literacy. In *Adaptive Educational Technologies for Literacy Instruction* (pp. 288–293).

Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA Educational Psychology Handbook, Vol 3: Application to Learning and Teaching* (pp. 451–473). Washington, DC: American Psychological Association.

Graesser, A. C., Hu, X., Rus, V., & Cai, Z. (2020). AutoTutor and other conversation-based learning and assessment environments. In A. Rupp, D. Yan, & P. Foltz (Eds.), *Handbook of Automated Scoring: Theory into Practice* (pp. 383–402). New York: CRC Press/Taylor and Francis.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 180–192.

Hayhoe, T., Podhorska, I., Siekelova, A., & Stehel, V. (2019). Sustainable manufacturing in Industry 4.0: Cross-sector networks of multiple supply chains, cyber-physical production systems, and AI-driven decision-making. *Journal of Self-Governance and Management Economics*, *7*(2), 31–36.

Holland, J., Mitrovic, A., & Martin, B (2009). J-LATTE: A constraint-based tutor for java. In S. C. Kong, H. Ogata, H. C. Arnseth, C. K. K. Chan, T. Hirashima, F. Klett, J. H. M. Lee, C. C. Liu, & C. K. Looi (Eds.), *Proceedings of 17th International Conference on Computers in Education ICCE 2009* (pp. 142–146). Hong Kong: Asia-Pacific Society for Computers in Education.

Hooshyar, D., Malva, L., Yang, Y., Pedaste, M., Wang, M., & Lim, H. (2021). An adaptive educational computer game: Effects on students' knowledge and learning attitude in computational thinking. *Computers in Human Behavior*, *114*, 106575.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, 105,* 1036–1049.

Jaques, N., Conati, C., Harley, J. M., & Azevedo, R. (2014). Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems* (pp. 29–38). Cham: Springer.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., … Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243.

Koedinger, K. R., & Corbett, A. (2006). *Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom.* The Cambridge Handbook of the Learning Sciences. New York, NY: Cambridge University Press.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42–78.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). *Handbook of latent Semantic Analysis*. Psychology Press.

Li, B. H., Hou, B. C., Yu, W. T., Lu, X. B., & Yang, C. W. (2017). Applications of artificial intelligence in intelligent manufacturing: A review. *Frontiers of Information Technology & Electronic Engineering*, *18*(1), 86–96.

Litman, D. J., Rosé, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, *16*(2), 145–170.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901.

McCarthy, K. S., Allen, L. K., & Hinze, S. R. (2020a). Predicting reading comprehension from constructed responses: Explanatory retrievals as stealth assessment. In *International Conference on Artificial Intelligence in Education* (pp. 197–202). Cham: Springer.

McCarthy, K. S., Watanabe, M., Dai, J., & McNamara, D. S. (2020b). Personalized learning in iSTART: Past modifications and future design. *Journal of Research on Technology in Education*, *52*(3), 301–321.

McNamara, D., Allen, L. K., McCarthy, S., & Balyan, R. (2018). NLP: Getting computers to understand discourse. In K. Millis, D. Long, J. Magliano, & K. Wiemer (Eds.), *Deep Learning: Multi-Disciplinary Approaches.* Routledge.

McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science, 2,* 1–15.

McNamara, D. S. (2021). Chasing theory with technology: A quest to understand understanding [Manuscript submitted for publication]. Department of Psychology, Arizona State University.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In *Handbook of latent Semantic Analysis* (pp. 227–241).

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499–515.

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., … Graesser, A. C. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution* (pp. 298–311). Hershey, PA: IGI Global.

Mitrovic, A. (2003). An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education*, *13*(2–4), 173–197.

Mitrovic, A., & Suraweera, P. (2016). Teaching database design with constraint-based tutors. *International Journal of Artificial Intelligence in Education*, *26*(1), 448–456.

Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*(1), 142–163.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24*(4), 427–469.

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*(5), 971–990.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127–144.

Paquette, L., & Baker, R. S. (2019). Comparing machine learning to knowledge engineering for student behavior modeling: A case study in gaming the system. *Interactive Learning Environments*, *27*(5–6), 585–597.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Francisco, CA: Morgan Kaufmann.

Peirce, N., Conlan, O., & Wade, V. (2008). Adaptive educational games: Providing non-invasive personalised learning experiences. In *2008 Second IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning* (pp. 28–35). Banff, Canada: IEEE.

Perrot, P. (2017). What about AI in criminal intelligence: From predictive policing to AI perspectives. *European Police Science and Research Bulletin*, *16*, 65–76.

Pham, P., & Wang, J. (2018). Adaptive review for mobile MOOC learning via multimodal physiological signal sensing-a longitudinal study. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 63–72). https://doi.org/10.1145/3242969.3243002.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322.

Price, T. W., Dong, Y., & Lipovac, D. (2017). iSnap: Towards intelligent tutoring in novice programming environments. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (pp. 483–488). New York, NY: ACM.

Psotka, J., Massey, L. D., & Mutter, S. A. (1988). *Intelligent Tutoring Systems: Lessons Learned*. Hillsdale: Lawrence Erlbaum Associates.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249–255.

Robertson, D. J., Noyes, E., Dowsett, A., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PLoS One, 11*(2), e0150036.

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59.

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, *105*(4), 1010.

Rus, V., D'Mello, S., Hu, X., & Graesser, A. C. (2013). Recent advances in intelligent systems with conversational dialogue. *AI Magazine*, *34*, 42–54.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Schultz, M. T. (2013). The intellimetric automated essay scoring engine-a review and an application to Chinese essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Scoring: Current Applications and Future Directions* (pp. 89–98). New York, NY: Routledge.

Sharma, K., Papamitsiou, Z., Olsen, J. K., & Giannakos, M. (2020). Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 480–489).

Shukor, N. A., & Abdullah, Z. (2019). Using learning analytics to improve MOOC instructional design. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(24), 6–17.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, *55*(2), 503–524.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present and future. In D. Jonassen (Ed.), *Handbook of Research on Educational Communications and Technology.* Scholastic Publications.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior, 116*, 106647. https://doi.org/10.1016/j.chb.2020.106647.

Shute, V., & Ventura, M. (2013). *Stealth Assessment: Measuring and Supporting Learning in Video Games.* Cambridge, MA: MIT Press.

Shute, V. J., Smith, G., Kuba, R., Dai, C. P., Rahimi, S., Liu, Z., & Almond, R. (2020). The design, development, and testing of learning supports for the physics playground game. *International Journal of Artificial Intelligence in Education*, 1–23.

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117.

Singh, V., & Gu, N. (2012). Towards an integrated generative design framework. *Design Studies*, *33*(2), 185–207.

Snow, E. L., Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education, 26*, 378–392.

Snow, E. L., Jackson, G. T., & McNamara, D. S. (2014). Emergent behaviors in computer-based learning environments: Computational signals of catching up. *Computers in Human Behavior, 41*, 62–70.

Snow, E. L., Likens, A. D., Allen, L. K., & McNamara, D. S. (2016). Taking control: Stealth assessment of deterministic behaviors within a game-based system. *International Journal of Artificial Intelligence in Education*, *26*, 1011–1032.

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105*(4), 970–987.

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal*, *18*(1), 87.

Taub, M., Azevedo, R., Rajendran, R., Cloude, E. B., Biswas, G., & Price, M. J. (2021). How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system? *Learning and Instruction*, *72*, 101200.

Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: Implications for recruitment. *Strategic HR Review, 17*(5), 255–258.

Van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215–222.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*, 227–265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., … Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, *15*(3), 147–204.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior, 29*(6), 2568–2572.

Ventura, M., Shute, V., & Zhao, W. (2013). The relationship between video game use and a performance-based measure of persistence. *Computers & Education*, *60*(1), 52–58.

Wang, C. H., & Lin, H. C. K. (2018). Emotional design tutoring system based on multimodal affective computing techniques. *International Journal of Distance Education Technologies*, *16*(1), 103–117.

Wang, L., Shute, V., & Moore, G. R. (2015). Lessons learned and best practices of stealth assessment. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), 66–87.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*, 1–24.

Worsley, M. (2018). Multimodal learning analytics: Past, present and potential futures. In *Proceedings of 8th International Conference on Learning Analytics & Knowledge*, Sydney, Australia.

Worsley, M., & Blikstein, P. (2018). A multimodal analysis of making. *International Journal of Artificial Intelligence in Education*, *28*(3), 385–419.

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731.