Situating AI (and Big Data) in the Learning Sciences: Moving Toward Large-Scale Learning Sciences

Danielle S. McNamara[1], Tracy Arner[1], Reese Butterfuss[1], Debshila Basu Mallick[2], Andrew S. Lan[3], Rod D. Roscoe[1], Henry L. Roediger III[4], and Richard G. Baraniuk[2]

[1] Arizona State University
[2] Rice University
[3] University of Massachusetts Amherst
[4] Washington University in St. Louis

McNamara D. S., Arner, T., Butterfuss, R., Mallick, D. B., Lan, A.S., Roscoe, R. D., Roediger III, H. L., & Baraniuk, R. G. (2022). Situating AI (and big data) in the learning sciences: Moving toward large-scale learning sciences. In A. Alavi, & B. McLaren (Eds.*), Artificial intelligence in STEM education: The paradigmatic shifts in research, education, and technology*. CRC Press.

Acknowledgements: The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Digital Learning Network Grants R305N210041 to Arizona State University and R305N210064 to Rice University. Funding was also partially provided by IES grants R305A190050 and R305A190063 and Office of Naval Research, through Grants N00014-17-1-2300 and N00014-20-1-2623 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education or the Office of Naval Research.

I. Introduction

The learning sciences inherently involve interdisciplinary research with an overarching objective of advancing theories of learning and to inform the design and implementation of effective instructional methods and learning technologies. In these endeavors, learning sciences encompass diverse constructs, measures, processes, and outcomes pertaining to both learning, motivation, and social interactions. These complex goals are further influenced by a large array of factors stemming from the learning context, learning task, and the characteristics of the individual learners. Learning occurs within a multitude of interacting contextual factors spanning variations between schools, teachers, classrooms, peers, and available technologies. These contexts also differ widely in diverse factors such as the social support that students receive, instructor engagement, demographic and ideological diversity, as well as instructional design strategies and affordances offered by educational technologies (Anderson & Dron, 2011). The learners themselves vary across a host of fixed factors such as age, grade level, ethnicity, and cultural background, as well as malleable individual differences such as engagement, interests, learning strategies, reading skills, and prior knowledge (Cantor et al., 2019; Jonassen & Grabowski, 2012; Winne, 1996).

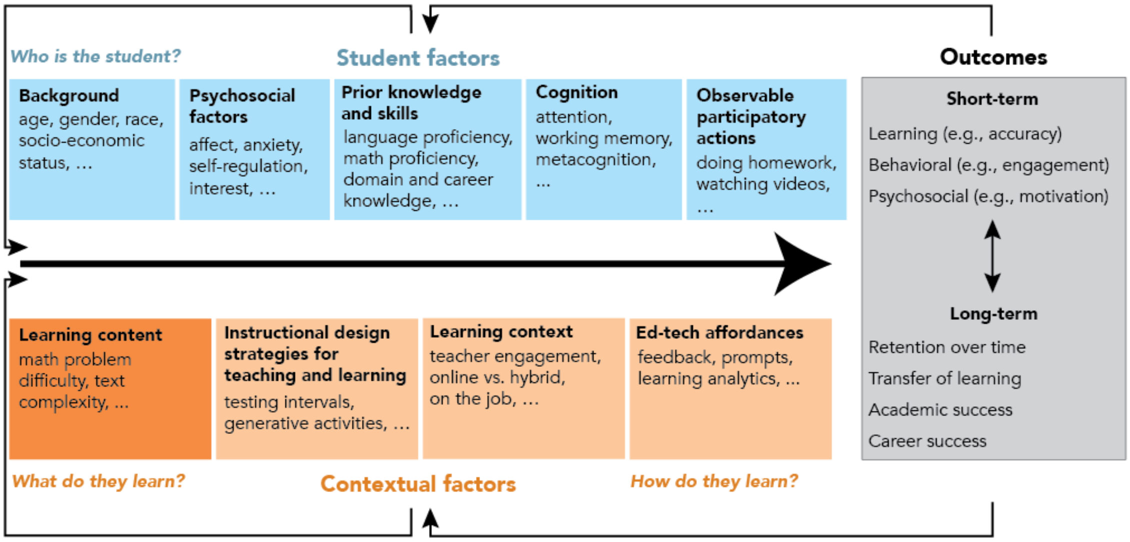

Who is the student?
Student factors
Outcomes
Background
age, gender, race, socio-economic status, ...
Psychosocial factors
affect, anxiety, self-regulation, interest, ...
Prior knowledge and skills
language proficiency, math proficiency, domain and career knowledge, ...
Cognition
attention, working memory, metacognition, ...
Observable participatory actions
doing homework, watching videos, ...
Short-term
Learning (e.g., accuracy)
Behavioral (e.g., engagement)
Psychosocial (e.g., motivation)
Long-term
Retention over time
Transfer of learning
Academic success
Career success
Learning content
math problem difficulty, text complexity, ...
Instructional design strategies for teaching and learning
testing intervals, generative activities, ...
Learning context
teacher engagement, online vs. hybrid, on the job, ...
Ed-tech affordances
feedback, prompts, learning analytics, ...
What do they learn?
Contextual factors
How do they learn?

Figure 1. *A Landscape of Learning Science Research***.** Numerous factors emerge from the characteristics of individual learners, learning contexts, and their interdependencies. Learners differ in fixed factors such as age, grade, ethnicity, and cultural background, as well as malleable individual differences such as engagement, interests, learning strategies, reading skills, and knowledge. Many studies that examine digital learning also assess observable participatory actions such as students' propensity to complete homework or attend class. In turn, the impact of individual differences depends on the learning context. Educational contexts differ widely in terms of instructor engagement, learner diversity, as well as instructional design strategies and affordances offered by educational technologies. The components included in each factor highlight only a subset of the variables that impact authentic student learning experiences.

Figure 1 visually conceptualizes this deeply complex landscape of learning science research with regard to student factors, contextual factors, and their interactive effects, with the ultimate objective of understanding how these factors impact short-term and long-term outcomes (e.g., grades, motivation, persistence, and test performance). A multitude of invaluable studies have examined these factors individually. Increasingly, however, researchers are diving into the complex interactions that emerge among individual differences (Yukselturk & Bulut, 2007) and then, in turn, investigating how the impact of such individual differences depends on the learning context (e.g., Coiro, 2021; Snow, 2002).

A large body of extant research already informs our understanding of how to improve learning across a variety of constrained contexts (e.g., laboratory experiments and targeted classroom studies). However, many of these assumptions, interventions, and findings have not yet been tested at scale or have failed to replicate beyond their initial, controlled environments. Moreover, existing work insufficiently draws from diverse theories and disciplines to examine the combined impact of multiple factors (Koedinger et al., 2013). Such narrow approaches are partially due to the combinatorial explosion that erupts when researchers study multiple complex factors and their relationships (Taber, 2019). In addition, many studies have been forced to assume or test linear relations between student and contextual factors, even though nonlinear relations are both theoretically and practically important. Finally, examining the impact of educational interventions in classrooms can be slow, as they inherently depend on academic timeframes spanning semesters or years.

Our supposition is that Artificial Intelligence (AI), naturally in combination with big data, provides multiple affordances with substantial potential to tackle the aforementioned challenges facing the learning sciences. AI--and even AIED--are sometimes viewed as antithetical to the learning sciences. In education, AI is often viewed as a tool to build applications and educational technologies that are designed based on learning science theories and evidence. Indeed, AI-based educational technologies have been used as a means to modify or enhance contextual factors. However, in contrast to this utilitarian perspective, AI is rarely viewed as a means to directly augment or enhance our understanding of learning itself or to advance the learning sciences. The purpose of this chapter is to situate AI and AIED within the learning sciences and to outline ways that AI can play a significant role in advancing the learning sciences.

In Section 2, we describe the landscape of learning science research (see Figure 1), including the broad range of individual and contextual factors and their relations to targeted educational outcomes as well as challenges that potentially inhibit significant advances in understanding and promoting learning. In response, Section 3 discusses affordances that arise from AI in this landscape and how AI can be leveraged to enhance, augment, and address challenges within the learning sciences. Finally, Section 4 proposes that leveraging AI in this manner has the potential to confront one of our biggest threats to education and society: inequitable systems that negatively impact the growth and success of learners and educators.

II. A Landscape of Learning Sciences and Some Challenges it Faces

2.1 Outcomes

One principal goal of education is to prepare students to become contributing members of a global knowledge economy (Dewey, 1934), and in turn, empower students to be agentic participants in pursuing their own goals and contributing to the betterment of society. In the learning sciences, attainment of these goals has been measured in a variety of ways, including standardized assessments (e.g., NAEP), completion of educational levels (e.g., high school, undergraduate degrees, and graduate degrees), course completion and grades, and demonstrated skill or knowledge development. For instance, students' retention of content and skills often serve as critical outcome measures of the effectiveness of contextual manipulations (i.e., instruction or intervention), wherein retention is typically operationalized as performance on multiple-choice tests or free responses to content questions (Hunt, 2003). Researchers can also examine how individual factors interact with contextual factors to identify for whom, and under what conditions, learning activities may be successful and thus contribute to desired outcomes of skill acquisition or knowledge retention.

Students' motivation also plays an integral role in the learning process -- students who engage more deeply with learning activities or persist through challenges show greater improvement in both short-term outcomes (e.g., retention; Alarcon & Edwards, 2013) and long-term outcomes (e.g., transfer; Cormier & Hagman, 2014; Haskell, 2000; Murayama et al., 2013). To assess motivation, researchers often collect self-report measures (Liu et al., 2012) to evaluate how students are internally and externally driven to complete learning tasks in varying contexts (Howard et al., 2021). Additionally, researchers have measured task engagement by tracking students' and teachers' behavioral indicators (e.g., time on task, participation, and communication patterns) and exploring how such behaviors relate to outcomes (e.g., Ocumpaugh et al., 2015). Lastly, measuring transfer of learning involves the design of learning sequences that allow researchers to track how earlier experiences and contexts influence later successes or struggles (Huang et al., 2009), as well as the learning strategies students use and their long-term effectiveness. Overall, the ultimate goals of much of learning sciences research are to understand how student factors and contextual factors influence different learning outcomes at different timescales.

2.2 Student Factors

Learning inherently depends on what the learner brings to the table. There are multiple aspects of individual learners that are correlated with educational outcomes, and potentially moderate or mediate the impact of instructional strategies or interventions, educational contexts, and affordances from technologies (see Figure 1; Cronbach & Snow, 1977; Preacher & Sterba, 2019). Investigations of student success have often considered conveniently accessible individual differences such as gender, race, and socioeconomic background (Wang, 2013). Students' academic history, including enrollment in Advanced Placement (AP) courses and performance indicators (i.e., GPA), can also be important for evaluating or predicting future success (Ma & Johnson, 2008; NAS, 2017). Student success and persistence can also be predicted by malleable factors such as general cognitive abilities, domain knowledge, literacy skills, along with motivational and social factors (e.g., engagement, perceived self-efficacy; Ackerman et al., 2013). Indeed, learners' prior knowledge is widely recognized as the single most important individual difference factor in education (Mayer, 2011; McCarthy & McNamara, 2021).

Although substantial research has examined the importance of individual differences (e.g., gender, motivation, intelligence, knowledge), less research has investigated how the effects of contextual factors (e.g., interventions) depend on those individual differences. Moreover, there are as many theories regarding individual differences as there are measures, and even the term “individual differences” evokes a wide range of constructs. Some theories focus on purportedly inherent abilities (e.g., working memory or general intelligence) and assume that students who perform well have more resources to process information (Alloway & Alloway, 2010; Cowan, 2014; Just & Carpenter, 1992). Other theorists focus on malleable skills such as reading skills and domain knowledge (Alexander et al., 1995; Perfetti, 2007), and assume that training such skills will improve students’ learning outcomes. Yet another focus is on differences in motivation, based on the theoretical assumption that intrinsically motivated students are more likely to perform well and succeed (Linnenbrink & Pintrich, 2002; Pintrich, 2003; Schunk & Zimmerman, 2012). In the context of instruction, one assumption is that *key* individual differences must be controlled for statistically. Alternatively, it is assumed that the effects of an instructional strategy or intervention might *depend* on a few key individual differences.

In the learning sciences, the question remains--how can the *right* individual differences be identified and measured to more comprehensively account for student learning? How can we match appropriate interventions to learners’ needs? One significant challenge in answering this question is that the field has advanced primarily by testing hypotheses in small-scale studies across narrow cross-sections of learners (Clarke & Dede, 2009; Dede, 2006; Kenny & Judd, 2019). Within these studies, limited sets and combinations of student and contextual variables are manipulated in tightly controlled laboratory studies (Makel & Plucker, 2014) because it is not possible to administer all of the measures that would cover the spectrum of meaningful individual differences. Thus, crucial individual differences (e.g., prior knowledge, literacy skills) that potentially influence the benefits of learning activities (McNamara, 2004, 2017; O’Reilly et al., 2004) are often not accounted for. A particularly high hurdle is obtaining a sufficiently large, representative, and diverse sample: a statistical interaction in a 2 x 2 factorial design requires approximately 4 times as many students as a simple main effect of the same magnitude (Fleiss, 1986). Without access to large pools of diverse students, it becomes impractical or impossible to answer the kinds of scientific questions that will drive personalized learning forward.

2.3 Contextual Factors

The impact of individual differences on learning outcomes is inherently influenced by the learning context. Contextual components of educational environments include what students learn (e.g., STEM domains, literacy skills), how students learn (e.g., instructional strategies), the instructional providers (e.g., instructors, pedagogical agents, intelligent tutoring systems), as well as affordances offered by educational technologies (e.g., precision education; Yang, 2021). For decades, learning science has investigated these components across multiple domains in an effort to reveal the best alignment between strategies, learners, and contexts (Dunlosky et al., 2013; Koedinger et al., 2013; Mayer, 2011). A number of robust effects have emerged from this work. The first of these is the generation effect (Bertsch et al., 2007). While the majority of generation effect studies have been restricted to episodic memory for familiar words (Gardiner, 1988; Slamecka & Katsaiti, 1987), unfamiliar words or phrases (Lutz et al., 2003), or answers to simple equations (McNamara & Healy, 1995a), generating has also been shown to enhance learning and skill acquisition (McNamara, 1995; McNamara & Healy, 1995b; Rittle-Johnson & Kmicikewycz, 2008).

The generation effect is comparable to the testing effect, wherein students attempt to retrieve content in a test-like format (e.g., cued recall or multiple choice), usually multiple times (Pyc & Rawson, 2009; Rawson & Dunlosky, 2012; Roediger & Butler, 2011). Compared to less active strategies (e.g., rereading text or notes), students benefit from practice testing even when retrieval is not successful (i.e., the answer is incorrect; Kornell et al., 2009). Relatedly, students also benefit from effortful retrieval when introducing space between study episodes (Bjork, 2014; cf. Soderstrom et al., 2015). During study sessions, students retrieve to-be-learned content which is activated in memory. Subsequent retrieval after some time lag (i.e., minutes, days, weeks) reactivates content in memory but allows for irrelevant details to fall away thus improving retention of target content (Cepeda et al., 2009, Vlach & Sandhofer, 2012). The benefit to students' subsequent recall varies with the type of content and the amount of time (e.g., minutes, days, weeks or months) between study sessions such that longer gaps between retrievals are more beneficial to long term retention (Rohrer, 2015). While the effect size varies, this effect is robust across grade levels and domains (Cepeda et al., 2009; Hintzman, 1974).

While various manipulations such as generating and repeated testing are helpful in enhancing memory, they are less helpful when the student is unable to understand the content. Often, students are challenged by the complexity of the content and the difficulty of the text. Self-explanation (explaininng text while reading) combined with reading comprehension strategies (e.g., paraphrasing, generating inferences) helps students to better comprehend challenging, unfamiliar content (McNamara, 2004, 2017). Producing self-explanations prompts students to make inferences between sentences in the text (i.e., bridging inferences) or between the text and prior knowledge (i.e., elaborative inferences).

Learning scientists generally agree that learning should not be passive. Learning contexts should include active, constructive activities (Chi & Wylie, 2014; Ebert-May et al., 1997; Mayer, 2009; Prince, 2004). Active learning has been operationalized differently across several disciplines and contexts (e.g., engineering, mathematics, medical education, science education, and engineering; Chamberland & Mamede, 2015; Crouch & Mazur, 2001; Freeman et al., 2014; Prince, 2004). Nonetheless, across studies, learning is generally enhanced to the degree that students engage in effortful, generative, constructive, participatory, and social learning activities (Fiorella & Mayer, 2016; Trafton & Trickett, 2001; Wittrock, 1989). The benefits of active learning have been evaluated in multiple contexts including interactive peer learning (Mazur, 1997), problem-based or inquiry-based learning (Hung et al., 2008), team-based learning (Sisk, 2011), collaborative learning (Menekse et al., 2013), and peer tutoring (Roscoe, 2014; Roscoe & Chi, 2008). Active learning, depending on its definition, is consistently found to be a key element for student success in developing essential 21st century skills such as collaboration and inquiry (Buitrago-Flórez et al., 2021; Christensen & Knezek, 2015).

The manner or mode in which instruction and feedback are delivered adds additional layers of complexity beyond the interactions between individual differences, content, and strategies. For example, the type (e.g., correct/incorrect or elaborative) and timing of feedback (e.g., immediate or delayed) have different effects based on the content type and the skill level of the learner (Fyfe & Rittle-Johnson, 2016; Koedinger et al., 2013; Kulik & Kulik; 1988). One key challenge is to identify for whom and under what conditions various instructional techniques are most effective for developing students' skills. Learning science researchers currently lack efficient tools for measuring the effectiveness of active learning at sufficiently large scales, which is needed to identify methods and conditions that maximize effectiveness. Moreover, learning sciences research is often conducted with internal constraints (e.g., student factors, teacher training, intervention fidelity) and external constraints (e.g., administrative support, cost, materials) imposed by limited funding, limited personnel, and a lack of infrastructure to support collaboration (Sabelli & Dede, 2013). Addressing these challenges requires valid and reliable measures of active learning to collect large-scale data from students' engagement and learning with online learning environments (Bryan et al., 2021).

2.4 Replication Crisis? (or Maybe Context Matters)

Unfortunately, studies demonstrating the impact of instructional practice often fail to demonstrate effectiveness beyond their initial learner populations, contexts, and scales (Bird et al., 2019; Dobronyi et al., 2019; Kizilcec et al., 2020; Lortie-Forgues & Inglis, 2019; Oreopoulos & Petronijevic, 2019; Yeager et al., 2019). This *replication crisis* has hit a number of fields, with many experiencing the pains of unreplicable findings far more sorely than the learning sciences.

Many of the challenges faced in the learning sciences are consequences of being restricted to small-scale studies, often in environments where behaviors and actions are untrackable. For example, randomized controlled trials (RCTs) have been a particularly popular assessment of efficacy; yet, Makel and Plucker (2014) found a paucity of published replications in the learning sciences. Given the complexity of bringing learning sciences innovations to scale, it is not surprising that replication attempts fail. Albeit not an exhaustive list, some reasons for such failures have included lack of fidelity (Nelson et al., 2012), questionable intervention design (Rodgers, 2016), and the assumption that students in authentic educational contexts will demonstrate benefits similar to those found in the context of laboratory settings (Clarke & Dede, 2009; Walker, 2004).

Another explanation for the absence of replication and generalization stems from the overarching assumption that ***individual differences matter***, and the impact of context depends on the learner. Inherent to learning sciences is that learning contexts are complex and organic, and that individuals vary in what they bring to the table, and what they need. While the small-scale studies approach makes sense, the multivariate and multidimensional nature of learning creates a combinatorial explosion that traditional experimental approaches just cannot accommodate. Thus, it has constrained our understanding of the complex interdependencies of students' individual differences and contextual factors that influence learning (Toh et al., 2016). Traditional studies do not yield sufficient data to inform the impact of various learning activities and interventions across diverse learner profiles and contexts (Bryan et al., 2021). This lack of data reflects a serious shortcoming of traditional small-scale studies in the learning sciences.

Challenges of replication and scale-up of learning sciences research may be partially mitigated by reducing reliance on traditional RCTs. RCTs are extremely valuable experimental approaches in many situations - depending on the design and choice of control/comparison conditions, an RCT provides the cleanest means of detecting the impact of an intervention, if you expect an intervention to work the same way across most contexts and most individuals. It is the scientist's stethoscope. However, RCTs are also costly, time-consuming, and can lack generalizability due to variations across samples and contexts (Lortie-Forgues & Inglis, 2019). Perhaps the largest shortcoming of RCTs in the context of learning sciences is the focus on manipulated factors without consideration of natural (important) variations in contexts and individuals. RCTs follow the assumption that all factors other than those that are under examination must be controlled. As illustrated in Figure 2, the scientist considers the classroom to provide something like a sterile petri dish, wherein they can observe behaviors in nicely controlled classrooms, as a function of some experimental manipulation. But, classrooms are not petri dishes. And, students are not cells to be cultured. The notion of *controlling* classroom environments, as if we were randomly administering medication is, well, just a bit ludicrous at best.

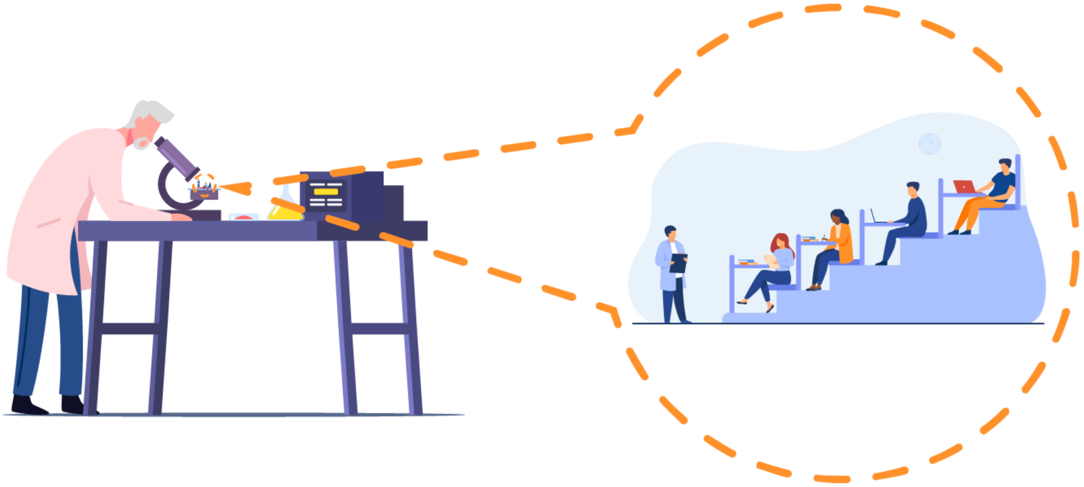


Figure 2. A Scientist Studies a *Well-Controlled* Classroom. Following the medical analogy, classrooms can be well controlled environments, like petri dishes, and students are identifiably pure, like cells to be cultured. However, classrooms must be organic and flexible, and students vary along a multitude of individual differences. This fundamental misconception regarding education reflects a potential shortcoming of relying so heavily on RCTs in the context of learning sciences (Illustration by Chris Kennedy).

Alternatively, resources may be better invested in developing infrastructure to support rapid, large-scale testing of interventions with diverse learners in a variety of authentic contexts (i.e., ***Large-Scale Learning Sciences*** or LSLS). Within industrial research (e.g., Google), rapid A/B (or A/B/n) experimentation is the process of simultaneously deploying variants of the same interface or design to different individuals and comparing which variant drives more activity or conversions. In doing so, A/B experimentation provides a wealth of data and is relatively cost-effective compared to implementing a series of traditional small-scale studies. Harnessing the power of big data has strong promise for addressing the shortcomings of small-scale learning sciences research. Namely, big data methodologies afford multiple, rapid, iterative, and cost-effective replications. Rapid A/B experiments also show promise for evaluating the effectiveness of interventions or educational systems across diverse learners and contexts (Lortie-Forgues & Inglis, 2019). This approach is particularly useful when testing the impact of variations in educational technologies and AI algorithms.

The past four decades of learning sciences research has greatly advanced our understanding of crucial learning outcomes, learners, and learning contexts (Ben-Eliyahu & Bernacki, 2015; Lee & Shute, 2010). Nonetheless, we have also observed that traditional learning sciences approaches cannot account for the breadth of student factors that influence learning. ***We do not fully understand learning in context***, principally because it is highly complex--perhaps too complex for existing approaches. Variations in context combined with the inherent complexities of learning and individual differences impose challenges in replicating, confirming, and applying the insights of these studies across different institutional contexts and broader demographic cross-sections of learners (Dede, 2006). Any combination of the variables discussed thus far can manifest within natural learning contexts, but they are difficult to detect, track, or measure.

III. AI and its Affordances for the Learning Sciences

Artificial Intelligence techniques are often leveraged in the context of educational technologies to provide information about individual factors (e.g., students' knowledge and skills), assess students' input as they engage with educational technologies, iteratively adapt contexts in attempts to modify students' performance, and in turn inform learning sciences (Corbett et al., 1997; Mousavinasab et al., 2021). Applications of machine learning in educational contexts often leverage educational datasets and use machine learning algorithms to build models that describe or predict student performance and other constructs of interest (Urbina Nájera & Calleja Mora, 2017). Advances in machine learning have enriched the design of computer-based learning environments because appropriate learning strategies and pedagogies provided to the learner can be derived specifically from attributes such as behaviors and performance within learning environments (Gamboa & Fred, 2002; Schiaffino et al., 2008). In turn, automated (intelligent) systems leverage AI to adjust task difficulty, learning paths, and provide feedback or support based on the information the systems acquire about students (Peirce et al., 2008; Van Eck, 2007).

One type of widely used educational technology that adapts the learning context based on student information is intelligent tutoring systems (ITSs). ITSs are computer programs that provide customized instruction and immediate feedback to learners (Psotka et al., 1988; VanLehn, 2006). Key features of ITSs include AI-driven feedback, real-time cognitive diagnosis and adaptive remediation (Shute & Psotka, 1996). When students interact with the ITS, the user interface presents learning materials to students through various media (e.g., pedagogical agents) and uses AI to interpret student input (i.e., speech, keystrokes, mouse clicks). Specifically, the student input data is processed and computationally analyzed, and then the analyzed input is used to update the student model and adjust the learning context to better suit the student. The "intelligent" aspects of systems like these are driven by AI to collect data about students' performance, conduct analyses, make instructional decisions to meet students' needs, and communicate with students (i.e., provide instruction or feedback; Shute & Psotka, 1996).

Opportunities to leverage AI in education have increased due to changes in the ways that instruction is delivered over the past two decades. We have seen dramatic shifts from a solely classroom-based instructional models to approaches that include some form of technology enhancement (i.e., hybrid) or to completely online instruction (e.g., synchronous, self-paced asynchronous; Clark & Mayer, 2011; Hamdan et al., 2013). This shift to technology-enhanced instruction provides new learning contexts and rich, personalized learning experiences (e.g., ITSs, virtual science labs, simulations) that were not previously available to students (Linn, 2004; Peffer et al., 2015; Slotta, 2010). Data science, data engineering, and machine learning provide powerful ways to collect and computationally analyze massive datasets and to untangle the complex interdependencies among large numbers of variables. The advent and proliferation of online learning in combination with advances in AI open up a wealth of opportunities, many of which are discussed in this volume of chapters. Here, we list five areas where we believe that AI (and big data) have strong potential to advance the learning sciences.

3.1 Deep Student Models

Student modeling (VanLehn, 1988) is an important research area as it provides key information on individual student factors that drive personalization. While a wide array of student models exists, we highlight two types to illustrate their importance. First, static item response theory models (van der Linden & Hambleton, 2013) analyze student responses to questions to estimate their knowledge levels on target knowledge components/skills/concepts. These models assume that student knowledge is static and are best suited for assessment purposes such as computerized, adaptive testing (Wainer et al., 2000). Second, knowledge tracing models (Corbett & Anderson, 1997) trace the evolution of student knowledge over time as they learn and are most effective for data collected over a longer time period.

The classic versions of these models are highly interpretable but cannot fully leverage large-scale student data made available recently, often consisting of tens of millions of student responses (Choi et al., 2020; Wang et al., 2021). Therefore, recent approaches have focused on using deep neural networks (Goodfellow et al., 2016) to improve the modeling capacity of student models (Piech et al., 2015; Wang et al., 2020). These methods achieve state-of-the-art performance in predicting next answer correctness but lose some interpretability, presenting challenges when deploying them in practice.

Therefore, future research on deep student models may benefit from considering the following two directions: First, integrating cognitive theory into large-scale, data-driven models to improve interpretability (Ghosh et al., 2020). Second, it is important to systematically study algorithmic biases, and in turn, devise more effective ways of detecting and mitigating biases in data (Gardner et al., 2019; Kizilcec & Lee, 2020). A paramount and growing concern regards the need to impose constraints during model training that promote fairness across different student groups and protect vulnerable participants (Yao & Huang, 2017).

3.2 Causal Learning Outcome Models

Most studies on student modeling focus on predictive models that are *correlational* in nature, based on observed student and contextual factors. However, in order to make discoveries regarding instructional strategies and content design that promote learning, we need *causal* models that directly relate student learning outcomes to changes in a contextual factor (Spirtes et al., 2000). Such models in turn will more effectively inform rigorous A/B experiments (de Carvalho et al., 2018; Sales & Pane, 2015). However, existing studies on causal modeling in education are limited by the scale as well as the structure and affordances of the datasets currently available (Sales et al., 2018).

As more and more platforms experiment with different instructional strategies and educational technology affordances, there is a possibility to obtain large-scale, quasi-experimental data that provide us with golden opportunities to develop these causal models (Gopalan et al., 2020). For example, one can use deep latent variable modeling approaches to learn both a representation of potential confounders and the strengths of causal effects between measured variables (Louizos et al., 2017). In the early stages of a study, when students experience a potential experimental condition, this type of analysis of quasi-experimental data will afford estimates of causal effect strengths between experimental variables and various individual factors. This can be achieved by including in the model both potential confounding variables and the estimated impact of the hypothesized causal factor. This requires that the wealth of existing data is mined to identify potential causal relations among various factors such that these factors can then be examined via subsequent, confirmatory rapid A/B experimentation.

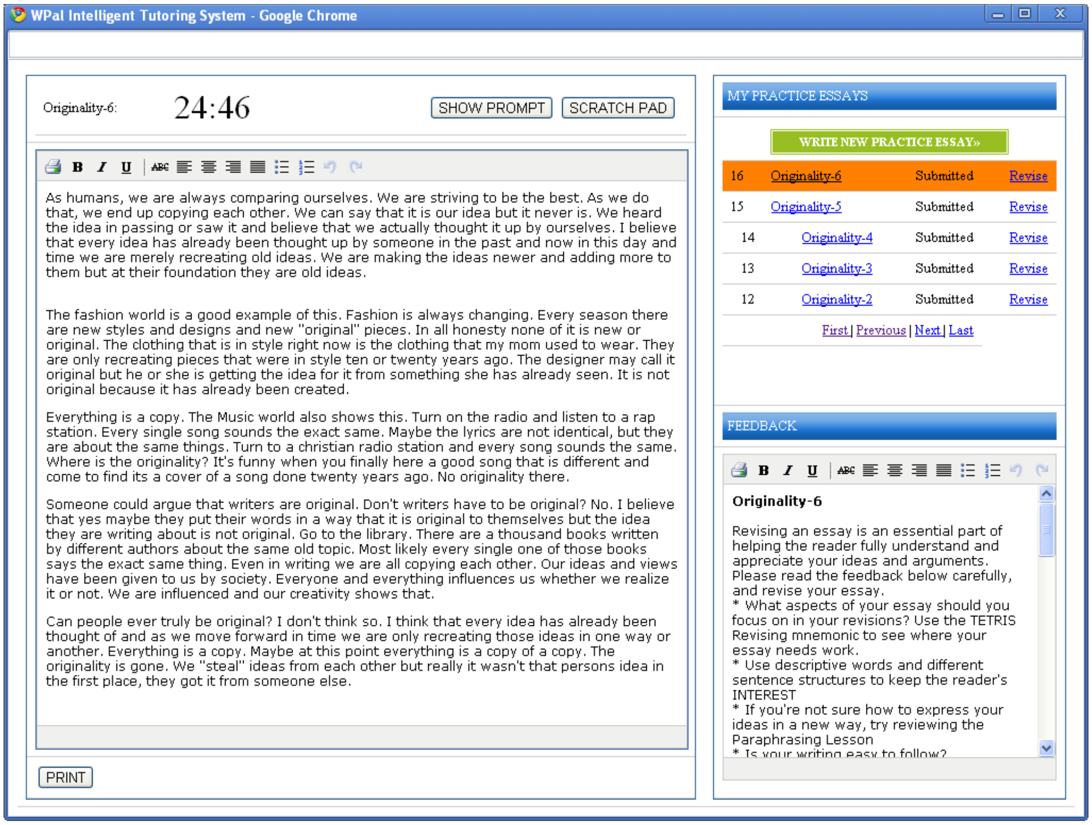
3.3 Natural Language Processing

Natural language processing (NLP) is the combination of computational linguistics and machine learning/AI to predict various aspects of the meaning or quality of the language (McNamara et al., 2017). Most learning environments include language, and for the most part, much of the rich interactions between students and among students goes unmined. Nonetheless, NLP has been used in various contexts to support student learning and assessment (Litman, 2016) and automatically score speech (Wang et al., 2018). The language that students use while learning is often key to understanding learning processes and predicting learning outcomes.

The most common use of NLP in the education domain is its use in automated scoring and assessment of writing quality (Burstein et al., 2017; Crossley et al., 2015). Automated essay scoring (AES) and automated writing evaluation (AWE) automatically assess students' essays and provide formative and summative feedback to writers during or after essay drafts (See Figure 3; Burstein, 2003; Foltz et al., 2013; Warschauer & Ware, 2006).

To develop these algorithms, machine learning is used to predict the quality of a corpus of essays that has been graded by experts based on linguistic and semantic features automatically extracted from the essays (McNamara et al., 2017). AES algorithms provide estimates of the summative scores related to the quality of essays (e.g., mechanics, content, cohesion, evidence). AWE algorithms, in turn, provide formative feedback that scaffold the writer in improving that essay, or future essays, with the objective of enhancing writing skills (Proske et al., 2014; Roscoe & McNamara, 2013; Roscoe et al., 2013; Wilson & Roscoe, 2020; Wilson et al., 2017).

NLP has also been leveraged to assess various aspects of learning materials, such as the difficulty of the readings and textbooks (McNamara et al., 2014). The most common approaches to assessing text difficulty (e.g., Flesch-Kincaid Grade Level) are based on features of the words and individual sentences, such as the length of word and sentence, or signals for word difficulty (e.g., word frequency) and sentence difficulty (e.g., complex syntax). Coh-Metrix was designed to go beyond word and sentence-level features to consider text cohesion (McNamara et al., 2014). Cohesion refers to the overlap in words, concepts, and ideas that impact the coherence and flow of the document. Cohesion is an important component of text difficulty because readers who are less knowledgeable about the topic have greater difficulties in filling in the cohesion gaps in text and consequently, are less likely to understand or learn from the reading materials (McNamara, 2017). AI comes into play here because it has been essential in developing the algorithms that combine the multitude of features within texts, and in turn predict whether they match the needs of the reader (Graesser & McNamara, 2011).


WPal Intelligent Tutoring System - Google Chrome
Originality-6:
24:46
SHOW PROMPT
SCRATCH PAD
PRINT
MY PRACTICE ESSAYS
WRITE NEW PRACTICE ESSAY»
FEEDBACK

As humans, we are always comparing ourselves. We are striving to be the best. As we do that, we end up copying each other. We can say that it is our idea but it never is. We heard the idea in passing or saw it and believe that we actually thought it up by ourselves. I believe that every idea has already been thought up by someone in the past and now in this day and time we are merely recreating old ideas. We are making the ideas newer and adding more to them but at their foundation they are old ideas.

The fashion world is a good example of this. Fashion is always changing. Every season there are new styles and designs and new "original" pieces. In all honesty none of it is new or original. The clothing that is in style right now is the clothing that my mom used to wear. They are only recreating pieces that were in style ten or twenty years ago. The designer may call it original but he or she is getting the idea for it from something she has already seen. It is not original because it has already been created.

Everything is a copy. The Music world also shows this. Turn on the radio and listen to a rap station. Every single song sounds the exact same. Maybe the lyrics are not identical, but they are about the same things. Turn to a christian radio station and every song sounds the same. Where is the originality? It's funny when you finally here a good song that is different and come to find its a cover of a song done twenty years ago. No originality there.

Someone could argue that writers are original. Don't writers have to be original? No. I believe that yes maybe they put their words in a way that it is original to themselves but the idea they are writing about is not original. Go to the library. There are a thousand books written by different authors about the same old topic. Most likely every single one of those books says the exact same thing. Even in writing we are all copying each other. Our ideas and views have been given to us by society. Everyone and everything influences us whether we realize it or not. We are influenced and our creativity shows that.

Can people ever truly be original? I don't think so. I think that every idea has already been thought of and as we move forward in time we are only recreating those ideas in one way or another. Everything is a copy. Maybe at this point everything is a copy of a copy. The originality is gone. We "steal" ideas from each other but really it wasn't that persons idea in the first place, they got it from someone else.

16	Originality-6	Submitted	Revise
15	Originality-5	Submitted	Revise
14	Originality-4	Submitted	Revise
13	Originality-3	Submitted	Revise
12	Originality-2	Submitted	Revise

First | Previous | Next | Last

Originality-6

Revising an essay is an essential part of helping the reader fully understand and appreciate your ideas and arguments. Please read the feedback below carefully, and revise your essay.
* What aspects of your essay should you focus on in your revisions? Use the TETRIS Revising mnemonic to see where your essay needs work.
* Use descriptive words and different sentence structures to keep the reader's INTEREST
* If you're not sure how to express your ideas in a new way, try reviewing the Paraphrasing Lesson
* Is your writing easy to follow?

Figure 3. *The Writing Pal Intelligent Tutoring System.* The Writing Pal automatically assesses students' essays using NLP-based algorithms and provides feedback to guide students' revisions as well as improve writing skills.

NLP can also be used examine specific aspects and attributes of the learner such as the relations of language sophistication, engagement, and collaboration on course performance (Crossley et al., 2015, 2016a, 2016b, 2017a, 2017b; Dascalu et al., 2018, 2020; Liu et al., 2016; San Pedro et al., 2015). Various NLP tools and methods can be used to extract dimensions and features of language within students' constructed responses (e.g., self-explanations, summaries, essays) and in turn infer various aspects of students' performance and individual differences (Allen & McNamara, 2015; McNamara et al., 2017). While approaches that leverage machine learning to combine multiple sources of information within educational contexts are promising, the bulk of this work has been limited to a narrow set of learning contexts. There is a need to develop infrastructure to help us examine learning beyond a few populations and learning contexts at a time.

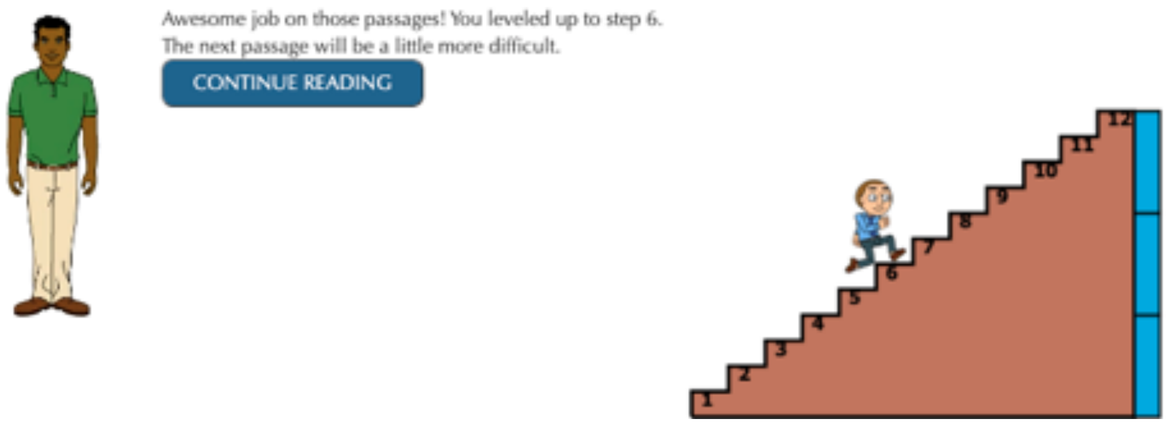
3.4 Sensor-free Student Factor Measures

To develop AI-based educational technologies, developers and educators need to make decisions about what information to collect about students in order to meet individual educational goals. AI offers the capability to measure student knowledge and skills covertly and unobtrusively using sensor-free measures. Many learning platforms have data on some student individual factors, such as fixed factors (e.g., grade level) and observables (e.g., past system interactions). However, we often do not have scalable measures of individual difference factors, such as motivation and social factors (e.g., engagement and affect) and ability factors (e.g., reading skill, prior knowledge). Traditional measures of these factors can be broadly categorized into two types. The first type resorts to student self-reports (Fredricks & McColskey, 2012), in-class observations (Ocumpaugh et al., 2015), or standardized assessments for individual factors such as reading skill. All of these require additional human effort and thus, are not immediately scalable. The second type uses physiological sensors that require external devices such as sensors (Suárez-Pellicioni et al., 2014) or cameras (Dragon et al., 2008), which can be expensive, invasive, and pose privacy concerns. In contrast, "sensor-free" detectors (e.g., embedded or stealth assessments) of these student characteristics gathered from activity and behavior logs have the potential to scale up measurement of these factors. After a relatively small number of self-reports, expert observations, or standardized assessments are administered to students to collect "ground truth" values, we can use AI and machine learning to predict these constructs using students' activity and behavior logs in online learning environments.

Stealth assessments are sensor-free assessments that are embedded in digital games (Shute, 2011; Shute & Ventura, 2013). In stealth assessment, student evaluations are embedded within gaming tasks and activities such that learners are unaware of being evaluated and they are not interrupted with overt tests or quizzes. Games inherently include rich sequences of actions; students' actions and performance during the gameplay are used to predict various aspects of students' knowledge and skills, as well as attributes like persistence and creativity (Shute et al., 2013, 2016; Shute & Rahimi, 2021; Ventura & Shute, 2013).

NLP can also be leveraged to provide sensor-free assessments of knowledge and skills (Allen & McNamara, 2015; Yang et al., 2019). For example, reading skill as measured by the Gates MacGinitie (MacGinitie & MacGinitie, 2006) can be predicted using linguistic indices extracted from self-explanations (e.g., Allen et al., 2015; McCarthy et al., 2020) and essays (e.g., Allen et al., 2016). Skilled readers are more likely to generate self-explanations containing greater semantic overlap and more explicit connections between their self-explanations (Allen et al., 2015). Additionally, readers who are prompted to self-explain produce more diverse language and global connectives in their responses, providing evidence that the cohesion of these constructed responses can serve as a proxy for coherence-building processes during comprehension (Allen et al., 2016; Flynn et al., in press). Fang et al. (2021) recently examined the number of student-generated self-explanations needed to predict reading comprehension skills during iSTART training (McCarthy et al., 2018, 2020). They found that the power of the linguistic features of self-explanations to predict reading comprehension skill increased as more self-explanations were included in the model, but only nine self-explanations were needed to explain 21% of the variance in reading comprehension skill. Ultimately, the objective is to combine *multiple* stealth measures of literacy skills (each separately accounting for different sources of variance) to provide a comprehensive profile of literacy, without the burden of administering standardized tests (McNamara et al., in press).

A principal benefit of AI-based assessments of individual factors is the ability to provide more precise individualized learning experiences for students. For example, AI-based stealth assessment of students' literacy skills in the context of an intelligent tutoring system can increase the precision of the system's adaptive features. Consequently, the system can provide automated literacy instruction and practice opportunities that are better suited to each student's skill profile (see Figure 4).



Awesome job on those passages! You leveled up to step 6.
The next passage will be a little more difficult.

Figure 4. *StairStepper Game in iSTART Intelligent Tutoring System***.** The StairStepper game in the iSTART intelligent tutoring system is an example of an AI-based stealth assessment of students' literacy skill level. NLP-based algorithms assess students' self-explanations and either increase or decrease the difficulty of the next text based on the quality of their self-explanation.

Additionally, AI-driven system analytics can provide early detection for students who are struggling before they are severely behind (Lu et al., 2018). In turn, teachers can use information educational technologies to inform instruction they provide in the classroom, potentially promoting more efficient use of instructional time and teachers' unique expertise. The wealth of data collected in AI-driven educational technologies can, and should, be leveraged to generate more informative learning analytics for teachers, thereby providing them with readily available, and more comprehensive windows into their students' current skill levels and progress.

3.5 Instructional Policy Learning

Instructional policies can contain typical low-level instructional planning (e.g., which learning a student should study at a point in time given their past learning record; Lan & Baraniuk, 2016) and high-level instructional strategies such as scaffolded support (Puntambekar & Kolodner, 2005), using spacing in retrieval practice (Roediger & Pyc, 2012), generative learning strategies (Mayer & Fiorella, 2015), monitoring and coping with student anxiety (Carver & Scheier, 1989; Huang & Mayer, 2016), and utility-value interventions (Canning & Harackiewicz, 2015). Learning instructional policies from student data will reveal what particular instructional or curricular strategy should be used, in what sequence, and to what depth. Thus, constructing personalized instructional policies based on each student's individual student factors by changing their contextual factors can maximize students' learning outcomes.

Learning an instructional policy can typically be viewed as a decision-making problem; reinforcement learning (RL) has been widely adopted for such problems in many applications from robot path planning to learning to play games where an agent can learn by interacting with the environment (Sutton & Barto, 2018). However, its application in real-world educational settings faces practical challenges, especially the problem of insufficient training data (see Doroudi et al., 2019, for an overview). Researchers have developed several ways to work around this issue and have found some success. First, one can reduce the complexity of the problem by restricting the set of instructional actions from which to select, usually down to two or three distinct actions (Zhou et al., 2019). Second, one can use student models learned from real student data to synthetically generate unlimited amounts of training data for RL algorithms (Choffin et al., 2020). Additionally, one can use other less data-hungry decision-making algorithms such as bandits (Lan & Baraniuk, 2016; Segal et al., 2018) and cognitive model-based algorithms in certain domains involving mainly memorization tasks (Upadhyay et al., 2021).

IV. Promoting Equity

To conduct valid research on understanding and improving learning requires scientists confront the (in)equitable and (un)ethical realities of educational contexts. Participants in learning environments pursue diverse and personal goals, have access to different capabilities and resources, and represent a variety of backgrounds and identities. Consequently, providing equitable and effective learning contexts is a persistent challenge. Successful learning requires culturally responsive approaches that recognize student agency and draw upon assets in a local context. Within the realm of AI and education, challenges of responsible computing are front and center as learning applications need to take extra care to be unbiased, transparent, and equitable. Indeed, most learning scientists are concerned with potential inequalities and inequities that data-driven AI methods may introduce. Many well-meaning interventions specifically address students who are most likely to struggle or be left behind. Yet, learning technologies are often less effective for historically underrepresented learners (Mayfield et al., 2019), and algorithms to predict and guide performance are often biased -- perpetuating and even magnifying inequities (Baker & Hawn, 2021; Kizilcec & Lee, 2020; Perry & Turner Lee 2019). Capturing and centering the "average" learning experience neglects the individuality of learners, which can be deleterious when the average ignores important variance.

Studies investigating aptitude-by-treatment interactions (and individual differences research, more generally) suggest that targeting an assumed average often fails to meet the needs of any one individual (e.g., Connor & Morrison 2016). Small population (< 1000 students) research has limited ability to adjust for variation in individual factors. Hence, results reported based on population means can limit applicability to historically underrepresented populations. Indeed, such limited research may be a causal factor underlying such historical underrepresentation (i.e., peoples' identities and needs cannot be examined within the data, then the data excludes those people). If important individual or contextual factors vary systematically with individual fixed factors (e.g., gender, ethnicity, learning English as a second language, etc.), then substantial explanatory elements for a particular student population are likely to go unnoticed under conventional research paradigms.

The bulk of the research focusing on educational inequalities or inequities has examined one or two social identities (e.g., race or gender) within narrowly defined contexts. For example, various studies have explored questions such as why and how gender gaps have evolved in STEM (Cheryan et al., 2017; Cimpian et al., 2020; Kanny et al., 2014; Soylu Yalcinkaya & Adams, 2020), why so few women major in Physics (Lewis et al., 2016; Walton et al., 2015), why Black and Hispanic students appear to achieve lower grades in middle school and beyond (Cohen et al., 2006), why STEM persistence rates differ by race/ethnicity (Riegle-Crumb et al., 2019), and the extent of such differences in online learning (Joosten & Cusatis, 2020; Mead et al., 2020; Wladis et al., 2016). These studies have yielded promising explanations and thus represent reasonable approaches when the study data are limited in size and scope.

By contrast, the use of AI and big data methods afford researchers the capacity to conduct research that examines multiple social identities simultaneously—including intersectional identities (e.g., Else-Quest & Hyde, 2016a, 2016b)—and explore the consequences to student outcomes and experiences longitudinally and comprehensively from their first day in college to graduation. We anticipate research that is able to combine multiple data sets and leverage AI to derive inferences across multiple populations and contexts will contribute to theories of human development, social support, self-directed and social learning, and computer-based learning in context, to name but a few.

To meet the needs of a broader array of students, researchers must conduct research with more diverse and inclusive samples of students and gather larger data sets that enable studying moderated and mediated relations across interventions and individual differences. In addition to diverse data sets, it is essential to collect *context-specific* data sets such that systems can be more responsive to a greater range of users who may have different needs and experiences (e.g., Bryan et al., 2021; Dolan, 2016). We must move beyond the petri dish analogy.

Issues of diversity, equity, and inclusion can no longer be an afterthought, instead they must be a critical consideration at every step of exploration, design, development, and implementation. Further, equity and inclusion is not limited to the technologies themselves. There must also be greater diversity across those involved in the design, development, and evaluation of these systems. AI and big data have the potential to help learning scientists more thoroughly understand learning and improve outcomes for *all* learners when diversity, equity, and inclusion are consistently included throughout all phases of learning sciences research. Working toward making educational data readily available to more researchers and providing instruction and training on how to leverage and interpret those data is vital to equity and diversity, and to developing a more comprehensive understanding of learning.

V. Conclusion

In the context of education, data science methods have provided important insights particularly through the emergence of educational data mining (Baker & Siemens, 2014; Romero & Ventura, 2010) and learning analytics (Lang et al., 2017) methods. Data science and data engineering provide powerful new ways to collect, curate, and computationally analyze massive datasets, and to disentangle the complex interdependencies among large numbers of variables. Modern educational technologies also facilitate the creation of engaging activities that may be further enhanced by AI and data science to adapt to students' specific needs to more effectively support learning, a practice that is not possible with traditional approaches to education (Nagar et al., 2019; Rau et al., 2017). However, as exciting as this possibility appears, adapting instruction to an individual in any sort of meaningful way is difficult because the existing science is not sufficiently advanced to provide guidance in all situations.

AI has a great deal to offer the learning sciences. However, without deep domain knowledge and relevant disciplinary expertise, it is often difficult to explain and interpret the relationships revealed by data science. Indeed, data science is typically agnostic about causal interpretations, which depend on theoretical frameworks outside of data science. AI (and data science) can reveal nuanced patterns of student retention, persistence, and performance related to demographic variables, but expertise in learning theory and psychological sciences is needed to suggest mechanisms and explanations for these patterns. Merging these domains and the knowledge and techniques therein is crucial to understanding learning.

References

Ackerman, P. L., Kanfer, R., & Beier, M. E. (2013). Trait complex, cognitive ability, and domain knowledge predictors of baccalaureate success, STEM persistence, and gender differences. *Journal of Educational Psychology, 105*(3), 911–927.

Alarcon, G. M., & Edwards, J. M. (2013). Ability and motivation: Assessing individual factors that contribute to university retention. *Journal of Educational Psychology, 105*(1), 129–137.

Alexander, P. A., Kulikowich, J. M., & Jetton, T. L. (1995). Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology, 87*(4), 559–575.

Allen, L. K., Dascalu, M., McNamara, D. S., Crossley, S., & Trausan-Matu, S. (2016). Modeling individual differences among writers using ReaderBench. *Proceedings of the 8th International conference on education and new learning technologies (EDULearn16)* (pp. 5269–5279). IATED.

Allen, L. K., & McNamara, D. S. (2015). Promoting self-regulated learning in an intelligent tutoring system for writing. In A. Mitrovic, F. Verdejo, C. Conati, & N. Heffernan (Eds.), *Doctoral consortium within the proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015)*, (pp. 827-830). Springer.

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 (LAK'15)* (pp. 246-254). Poughkeepsie, NY: ACM.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20-29.

Anderson, T., & Dron, J. (2011). Three generations of distance education pedagogy. *International Review of Research in Open and Distributed Learning*, *12*(3), 80-97.

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. *edarxiv.org*

Baker, R.S., & Siemens, G. (2014). Educational data mining and learning analytics. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 253-274). Cambridge University Press.

Ben-Eliyahu, A., & Bernacki, M. L. (2015). Addressing complexities in self-regulated learning: A focus on contextual factors, contingencies, and dynamic relations. *Metacognition and Learning*, *10*(1), 1-13.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201-210.

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2019). Nudging at scale: Experimental evidence from FAFSA completion campaigns. NBER Working Paper No. 26158. *National Bureau of Economic Research*.

Bjork, R.A. (2014). Forgetting as a friend of learning. In D.S. Lindsay, C. M. Kelley, A.P. Yonelinas, & H.l. Roediger (Eds.), *Remembering: Attributions, processes, and control in human memory-Papers in honour of Larry L. Jacoby* (pp. 15-28). Psychology Press.

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, 1-10.

Buitrago-Flórez, F., Danies, G., Restrepo, S., & Hernández, C. (2021). Fostering 21st Century Competences through Computational Thinking and Active Learning: A Mixed Method Study. *International Journal of Instruction*, *14*(3).

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion Online essay evaluation: an application for automated evaluation of student essays. *Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence* (pp. 3-10).

Burstein, J., McCaffrey, D., Beigman Klebanov, B., & Ling, G. (2017). Exploring relationships between writing and broader outcomes with automated writing evaluation. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.,), *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. Copenhagen, Denmark: Association for Computational Linguistics.

Canning, E. A., & Harackiewicz, J. M. (2015). Teach it, don't preach it: The differential effects of directly-communicated and self-generated utility–value information. *Motivation Science, 1*(1), 47.

Cantor, P., Osher, D., Berg, J., Steyer, L., & Rose, T. (2019). Malleability, plasticity, and individuality: How children learn and develop in context. *Applied Developmental Science*, *23*(4), 307-337.

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: a theoretically based approach. *Journal of Personality and Social Psychology, 56*(2), 267.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236-246.

Chamberland, M., & Mamede, S. (2015). Self-explanation, an instructional strategy to foster clinical reasoning in medical students. *Health Professions Education*, *1*(1), 24-33.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1-35.

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219-243.

Choffin, B., Popineau, F., & Bourda, Y. (2020). Modelling student learning and forgetting for optimally scheduling skill review. *ERCIM News*, *2020*(120), 12-13.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., ... & Heo, J. (2020). Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education* (pp. 69-73). Springer, Cham.

Christensen, R., & Knezek, G. (2015). Active learning approaches to integrating technology into a middle school science curriculum based on 21st century skills. In *Emerging Technologies for STEAM Education* (pp. 17-37). Springer.

Cimpian, J. R., Kim, T. H., & McDermott, Z. T. (2020). Understanding persistent gender gaps in STEM. *Science*, *368*(6497), 1317-1319.

Clark, R. C., & Mayer, R. E. (2011). *E-learning and the science of instruction proven guidelines for consumers and designers of multimedia learning, Third edition* (3rd ed.). Pfeiffer.

Clarke, J., & Dede, C. (2009). Robust designs for scalability. In L. Moller, J. B. Huett, D. M. Harvey (Eds.), *Learning and instructional technologies for the 21st century*. Springer.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*(5791), 1307-1310.

Coiro, J. (2021). Toward a multifaceted heuristic of digital reading to inform assessment, research, practice, and policy. *Reading Research Quarterly, 56,* 9-31.

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: Implications for policy and practice. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 54-61.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, *4*(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.

Cormier, S. M., & Hagman, J. D. (Eds.). (2014). *Transfer of learning: Contemporary research and applications*. Academic Press.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, *26*(2), 197-223.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.

Crossley, S., Barnes, T., Lynch, C., & McNamara, D. S. (2017a). Linking language to math success in an on-line course. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining (EDM)* (pp. 180-185). International Educational Data Mining Society.

Crossley, S. A., Dascalu, M., Baker, M., McNamara, D. S., & Trausan-Matu, S. (2017b). Predicting success in massive open online courses (MOOC) using cohesion network analysis. In B. K. Smith, M. Borge, K. Y. Lim, & E. Mercier (Eds.), *Proceedings of the 12th International Conference on Computer-Supported Collaborative Learning* (CSCL 2017). (pp. 103-110). ISLS.

Crossley, S. A, Kyle, K., Davenport, J., & McNamara, D. S. (2016a). Automatic assessment of constructed response data in a chemistry tutor. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)* (pp. 336-340). International Educational Data Mining Society.

Crossley, S., McNamara, D. S., Baker, R., Wang, Y., Paquette, L, Barnes, T., & Bergner, Y. (2015). Language to completion: Success in an educational data mining massive open online class. 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. 388-391). International Educational Data Mining Society.

Crossley, S. A., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. (2016b). Combining click-stream data with NLP tools to better understand MOOC completion. In D. Gašević, G. Lynch, S. Dawson, H. Drachsler, & C. P. Rosé (Eds.), *Proceedings of the 6th International Learning Analytics & Knowledge Conference* (LAK'16), (pp. 6-14). ACM.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977.

Dascalu, M. D., Dascalu, M., Ruseti, S., Carabas, M., Trausan-Matu, S., & McNamara, D. S. (2020). Cohesion network analysis: Predicting course grades and generating sociograms for a Romanian Moodle course. *Proceedings of 16th International Conference on Intelligent Tutoring Systems* (pp. 174-183). Springer.

Dascalu, M., Sirbu, M. D., Gutu-Robu, G., Ruseti, S., Crossley, S. A., & Trausan-Matu, S. (2018). Cohesion-centered analysis of sociograms for online communities and courses using ReaderBench. *European Conference on Technology Enhanced Learning* (pp. 622-626). Springer.

de Carvalho, W. F., Couto, B. R. G. M., Ladeira, A. P., Gomes, O. V., & Zarate, L. E. (2018). Applying causal inference in educational data mining: A pilot study. *Proceedings of the 10th International Conference on Computer Supported Education (1)* (pp. 454-460).

Dede, C. (2006). Scaling up: Evolving innovations beyond ideal settings to challenging contexts of practice. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences*. Cambridge University Press.

Dewey, J. (1934). Individual psychology and education. *The Philosopher*, *12*(1), 1-6.

Dobronyi, C. R., Oreopoulos, P., & Petronijevic, U. (2019). Goal setting, academic reminders, and college success: A large-scale field experiment. *Journal of Research on Educational Effectiveness*, *12*(1), 38-66.

Dolan, J. E. (2016). Splicing the divide: A review of research on the evolving digital divide among K–12 students. *Journal of Research on Technology in Education*, *48*(1), 16-37.

Doroudi, S., Aleven, V., & Brunskill, E. (2019). Where's the reward?. *International Journal of Artificial Intelligence in Education*, *29*(4), 568-620.

Dragon, T., Arroyo, I., Woolf, B. P., Burleson, W., El Kaliouby, R., & Eydgahi, H. (2008). Viewing student affect and learning through classroom observation and physical sensors. *International Conference on Intelligent Tutoring Systems* (pp. 29-39). Springer.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.

Ebert-May, D., Brewer, C., & Allred, S. (1997). Innovation in large lectures: Teaching for active learning. *Bioscience*, *47*(9), 601-607.

Else-Quest, N. M., & Hyde, J. S. (2016a). Intersectionality in quantitative psychological research: I. Theoretical and epistemological issues. *Psychology of Women Quarterly*, *40*(2), 155-170.

Else-Quest, N. M., & Hyde, J. S. (2016b). Intersectionality in quantitative psychological research: II. Methods and techniques. *Psychology of Women Quarterly*, *40*(3), 319-336.

Fang, Y., Li, T., Roscoe, R. D., & McNamara, D. S. (2021). Predicting literacy skills via stealth assessment in a simple vocabulary game. *Proceedings of International Conference on Human-Computer Interaction* (pp. 32-44). Springer.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*(4), 717-741.

Fleiss, J. L. (1986). Reliability of measurement. In J. L. Fleiss (Ed.). *The design and analysis of clinical experiments* (pp. 1–32). John Wiley and Sons.

Flynn, L. E., McNamara, D. S., McCarthy, K. S., Magliano, J. P., & Allen L. K. (in press). The appearance of coherence: Using cohesive properties of readers' constructed responses to predict individual differences. *Revista Signos. Estudios de Lingüística.*

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 68–88). Routledge.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S.

Christenson, A. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763-782). Springer.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410-8415.

Fyfe, E. R., & Rittle-Johnson, B. (2016). Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology*, *108*(1), 82-97.

Gamboa, H., & Fred, A. (2002). Designing intelligent tutoring systems: A bayesian approach. In J. Filipe, B. Sharp, & P. Miranda (Eds.), *Enterprise Information Systems III* (pp. 146-152). Kluwer Academic Publishers.

Gardiner, J. M. (1988). Generation and priming effects in word-fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 495–501.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 225-234).

Ghosh, A., Heffernan, N., & Lan, A. S. (2020). Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2330-2339).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, *44*(1), 218-243.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*, 371–398.

Hamdan, N., McKnight, P., McKnight, K., & Arfstrom, K. (2013). A Review of Flipped Classroom. *Flipped Learning Network.*

Haskell, R. E. (2000). *Transfer of learning: Cognition and instruction*. Elsevier.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77-97). Lawrence Erlbaum Associates.

Howard, J. L., Bureau, J., Guay, F., Chong, J. X., & Ryan, R. M. (2021). Student motivation and associated outcomes: A meta-analysis from self-determination theory. *Perspectives on Psychological Science*, 1745691620966789.

Huang, Y. M., Huang, T. C., Wang, K. T., & Hwang, W. Y. (2009). A Markov-based recommendation model for exploring the transfer of learning on the web. *Journal of Educational Technology & Society*, *12*(2), 144-162.

Huang, X., & Mayer, R. E. (2016). Benefits of adding anxiety-reducing features to a computer-based multimedia lesson on statistics. *Computers in Human Behavior, 63*, 293-303.

Hung, W., Jonassen, DH., & Liu, R. (2008). Problem-Based Learning. In M. Spector, D. Merrill, J. van Merrienböer, M. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 485-506). Erlbaum.

Hunt, D. P. (2003). The concept of knowledge and how to measure it. *Journal of Intellectual Capital, 4*(1), 100-113.

Jonassen, D. H., & Grabowski, B. L. (2012). *Handbook of individual differences, learning, and instruction*. Routledge.

Joosten, T., & Cusatis, R. (2020). Online learning readiness. *American Journal of Distance Education*, *34*(3), 180-193.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122–149.

Kanny, M. A., Sax, L. J., & Riggers-Piehl, T. A. (2014). Investigating forty years of STEM research: How explanations for the gender gap have evolved over time. *Journal of Women and Minorities in Science and Engineering*, *20*(2).

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578.

Kizilcec, R. F., & Lee, H. (2020). Algorithmic fairness in education. *arXiv preprint arXiv:2007.05443*.

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkey, S., Williams, J. J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26), 14900-14905.

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. *Science*, *342*(6161), 935-937.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989-998.

Kulik, J. A., & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79-97.

Lan, A. S., & Baraniuk, R. G. (2016). A Contextual Bandits Framework for Personalized Learning Action Selection. In *EDM* (pp. 424-429).

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2017). *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research.

Lee, J., & Shute, V. J. (2010). Personal and social-contextual factors in K–12 academic performance: An integrative perspective on student learning. *Educational Psychologist, 45*(3), 185-202.

Lewis, K. L., Stout, J. G., Pollock, S. J., Finkelstein, N. D., & Ito, T. A. (2016). Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics. *Physical Review Physics Education Research*, *12*(2), 020110.

Linn, M. C., Davis, E. A., & Bell, P. (2004). Inquiry and technology. *Internet Environments for Science Education* (pp. 3–28).

Linnenbrink, E. A., & Pintrich, P. R. (2002). Motivation as an enabler for academic success. *School Psychology Review*, *31*(3), 313-327.

Litman, D. (2016). Natural language processing for enhancing teaching and learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 30*(1). https://ojs.aaai.org/index.php/AAAI/article/view/9879

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*(9), 352-362.

Liu, Z., Brown, R., Lynch, C.F., Barnes, T., Baker, R., Bergner, Y., & McNamara, D. S. (2016). MOOC learner behaviors by country and culture: An exploratory analysis. T. Barnes, M. Chi, & M. Feng (Eds.), In *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)* (pp.127-134). International Educational Data Mining Society.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned?. *Educational Researcher*, *48*(3), 158-166.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. In I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6449-6459). Curran Associates Inc.

Lu, O., Huang, A., Huang, J., Lin, A., Ogata, H., & Yang, S. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Journal of Educational Technology & Society, 21*(2), 220-232.

Lutz, J., Briggs, A., & Cain, K. (2003). An examination of the value of the generation effect for learning new material. *The Journal of General Psychology*, *130*(2), 171-188.

Ma, X., & Johnson, W. (2008). Mathematics as the critical filter: Curricular effects on gendered career choices. In H. M. G. Watt & J. S. Eccles (Eds.), *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences* (pp. 55–83). American Psychological Association.

MacGinitie, W. H., & MacGinitie, R. K. (2006). *Gates-MacGinitie reading tests* (4th ed.). Houghton Mifflin.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304-316.

Mayer, R. E. (2009). Constructivism as a theory of learning versus constructivism as a prescription for instruction. In *Constructivist instruction* (pp. 196-212). Routledge.

Mayer, R. E. (2011). *Applying the science of learning.* Pearson.

Mayer, R. E., & Fiorella, L. (2015). *Learning as a generative activity*. Cambridge, UK: Cambridge University Press.

Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E., & Black, A. W. (2019, August). Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 444-460).

Mazur, E. (1997). Peer instruction: Getting students to think in class. In *AIP Conference Proceedings* (Vol. 399, No. 1, pp. 981-988). AIP.

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

McCarthy, K. S., Likens, A. D., Kopp, K. J., Watanabe, M., Perret, C. A., & McNamara, D. S. (2018). The "LO"-down on grit: Non-cognitive trait assessments fail to predict learning gains in iSTART and W-Pal. In *Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK'18).*

McCarthy, K. S., & McNamara, D. S. (2021). The multidimensional knowledge in text comprehension framework. *Educational Psychologist.*

McNamara, D. S. (1995). Effects of prior knowledge on the generation advantage: Calculators versus calculation to learn simple multiplication. *Journal of Educational Psychology, 87,* 307-318.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*(1), 1-30.

McNamara, D. S. (2017). Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. *Discourse Processes*, *54*(7), 479-492.

McNamara, D. S., Allen, L. K., Crossley, S. A., Dascalu, M., & Perret, C. A. (2017). Natural language processing and learning analytics. In G. Siemens, & C. Lang (Eds.), *Handbook of learning analytics and educational data mining* (pp. 93-104). Society for Learning Analytics Research.

McNamara, D. S., Fang, Y., Butterfuss, M., Arner, T., Watanabe, M., McCarthy, K. S., Allen, L. K., & Roscoe, R. D. (in press). iSTART: Adaptive comprehension strategy training and stealth literacy assessment. *International Journal of Human-Computer Interaction.*

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014*). Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press.

McNamara, D. S., & Healy, A. F. (1995a). A procedural explanation of the generation effect: The use of an operand retrieval strategy for multiplication and addition problems. *Journal of Memory and Language, 34,* 399-416.

McNamara, D. S., & Healy, A. F. (1995b). A generation advantage for multiplication skill and nonword vocabulary acquisition. In A.F. Healy & L.E. Bourne, Jr. (Eds.), *Learning and memory of knowledge and skills* (pp. 132-169). Sage.

Mead, C., Supriya, K., Zheng, Y., Anbar, A. D., Collins, J. P., LePore, P., & Brownell, S. E. (2020). Online biology degree program broadens access for women, first-generation to college, and low-income students, but grade disparities remain. *PloS One*, *15*(12), e0243916.

Menekse, M., Stump, G. S., Krause, S., & Chi, M. T. (2013). Differentiated overt learning activities for effective instruction in engineering classrooms. *Journal of Engineering Education, 102*(3), 346-374.

Mousavinasab, E., Zarifsanaiey, N., R. 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.

Murayama, K., Pekrun, R., Lichtenfeld, S., & Vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development*, *84*(4), 1475-1490.

Nagar, N., Shachar, H. & Argaman, O. (2019). Changing the Learning Environment: Teachers and Students' Collaboration in Creating Digital Games. *Journal of Information Technology Education: Innovations in Practice, 18*(1), 61-85. Informing Science Institute. Retrieved August 13, 2021 from https://www.learntechlib.org/p/216644/.

National Academies of Sciences, Engineering, and Medicine. (2017). *The economic and fiscal consequences of immigration.* National Academies Press. https://www.jstor.org/stable/pdf/44202635.pdf?casa_token=NNMEle4_gSkAAAAA:7ew66pDIAulpMdPVdG205XFaRBlwLwcPhBCKSUZwiwbjeIpy1RmGXZjgxRJNT0HwhI7_L3iOpUokQ9sq7yetVuanYAwFjX-QGDlc1cRyyP25wvoWb9nh

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services & Research, 39*(4), 374-396.

Ocumpaugh, J., Baker, R. S., & Rodrigo, M. M. T. (2015). *Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual.* Columbia University, Manila University.

O'Reilly, T., Best, R., & McNamara, D. S. (2004). Self-explanation reading training: Effects for low-knowledge readers. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 26, No. 26).

Oreopoulos, P., & Petronijevic, U. (2019). *The remarkable unresponsiveness of college students to nudging and what we can learn from it*. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w26059/w26059.pdf

Peffer, M.E., Beckler, M.L., Schunn, C., Renken, M., & Revak, A. (2015) Science classroom inquiry (SCI) simulations: A novel method to scaffold science learning. *PLoS ONE 10*(3): e0120638.

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). IEEE.

Perfetti, C.A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357–383.

Perry, A. M. & Turner Lee, N (2019). *AI is coming to schools, and if we're not careful, so will its biases.* https://www.brookings.edu/blog/the-avenue/2019/09/26/ai-is-coming-to-schools-and-if-w ere-not-careful-so-will-its-biases/

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *arXiv preprint arXiv:1506.05908*.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95,* 667-686.

Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-treatment interactions in research on educational interventions. *Exceptional Children, 85*(2), 248-264.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*(3), 223-231.

Proske, A., Roscoe, R. D., & McNamara, D. S. (2014). Game-based practice versus traditional practice in computer-based writing strategy training: Effects on motivation and achievement. *Educational Technology Research and Development*, *62*(5), 481-505.

Psotka, J., Massey, L. D., & Mutter, S. A. (1988). *Intelligent tutoring systems: Lessons learned.* Lawrence Erlbaum Associates.

Puntambekar, S., & Kolodner, J. L. (2005). Toward implementing distributed scaffolding: Helping students learn science from design. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 42*(2), 185-217.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437-447.

Rau, M. A., Kennedy, K., Oxtoby, L., Bollom, M., & Moore, J. W. (2017). Unpacking "active learning": A combination of flipped classroom and collaboration support is more effective but collaboration support alone is not. *Journal of Chemical Education*, *94*(10), 1406-1414.

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning?. *Educational Psychology Review*, *24*(3), 419-435.

Riegle-Crumb, C., King, B., & Irizarry, Y. (2019). Does STEM stand out? Examining racial/ethnic gaps in persistence across postsecondary fields. *Educational Researcher*, *48*(3), 133-144.

Rittle-Johnson, B., & Kmicikewycz, A. O. (2008). When generating answers benefits arithmetic skill: The importance of prior knowledge. *Journal of Experimental Child Psychology*, *101*(1), 75-81.

Rodgers, E. (2016). Scaling and sustaining an intervention: The case of Reading Recovery. *Journal of Education for Students Placed at Risk (JESPAR)*, *21*(1), 10-28.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27.

Roediger III, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248.

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review, 27*(4), 635-643.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618.

Roscoe, R. D. (2014). Self-monitoring and knowledge-building in learning by teaching. *Instructional Science, 42*(3), 327-351.

Roscoe, R. D., & Chi, M. T. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science, 36*(4), 321-350.

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.

Roscoe, R. D., Varner, L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology 25, 8*(4), 362-381.

Sabelli, N., & Dede, C. (2013). Empowering design based implementation research: The need for infrastructure. *National Society for the Study of Education, 112*(2), 464-480.

Sales, A., Botelho, A. F., Patikorn, T., & Heffernan, N. T. (2018). Using big data to sharpen design-based inference in A/B tests. In *Proceedings of the Eleventh International Conference on Educational Data Mining* (pp. 479-485). National Science Foundation.

Sales, A. C., & Pane, J. F. (2015). Exploring causal mechanisms in a randomized effectiveness trial of the cognitive tutor. In *Proceedings of the 8th International Conference on Educational Data Mining*.

San Pedro, M. O. Z., Snow, E. L., Baker, R. S., McNamara, D. S., & Heffernan, N. T. (2015). Exploring dynamical assessments of affect, behavior, and cognition and math state test achievement. 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. 85-92). International Educational Data Mining Society.

Schiaffino, S., Garcia, P., & Amandi, A. (2008). eTeacher: Providing personalized assistance to e-learning students. *Computers & Education*, *51*(4), 1744-1754.

Schunk, D., & Zimmerman, B. (2012). *Motivation and self-regulated learning: Theory, research, and applications*. Routledge.

Segal A., Ben David Y., Williams J.J., Gal K., Shalom Y. (2018) Combining Difficulty Ranking with Multi-Armed Bandits to Sequence Educational Content. In: Penstein Rosé C. et al. (eds) *Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science*, vol 10948. Springer, Cham.

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. 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.

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games* (p. 102). The MIT Press.

Shute, V., Ventura, M., Small, M., & Goldberg, B. (2013). Modeling student competencies in video games using stealth assessment. In R. Sottilare, A. Graesser, X. Hu & H. Holden (Eds.), *Design recommendations for intelligent tutoring systems*, Vol. 1: Learner Modeling (pp. 141–152).

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.

Sisk, R. J. (2011). Team-based learning: Systematic research review. *Journal of Nursing Education*, *50*(12), 665-669.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language*, *26*(6), 589-607.

Slotta, J. I. M. (2010). Evolving the classrooms of the future: The interplay of pedagogy, technology, and community. In *Classroom of the Future* (pp. 215-242). Brill Sense.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* Rand Education. https://apps.dtic.mil/sti/pdfs/ADA402712.pdf

Soderstrom, N., & Kerr, T., & Bjork, R. L. (2015). The critical importance of retrieval-and spacing-for learning. *Psychological Science, 27*(2), 223-230.

Soylu Yalcinkaya, N., & Adams, G. (2020). A cultural psychological model of cross-national variation in gender gaps in STEM participation. *Personality and Social Psychology Review*, *24*(4), 345-370.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Suárez-Pellicioni, M., Núñez-Peña, M.I., & Colomé, À. (2014). Reactive recruitment of attentional control in math anxiety: An ERP study of numeric conflict monitoring and adaptation. *PLoS ONE 9*(6): e99579.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Taber, K. S. (2019). Experimental research into teaching innovations: responding to methodological and ethical challenges. *Studies in Science Education*, *55*(1), 69-119.

Toh, Y., Lee, J. Y. L., & Ting, K. S. W. (2016). Building synergies: Taking school-based interventions to scale. In C. S. Chai, C. P. Lim, & C. M. Tan (Eds.), *Future learning in primary schools: A Singapore perspective* (pp. 177-197). Springer.

Trafton, J. G., & Trickett, S. B. (2001). Note-taking for self-explanation and problem solving. *Human-Computer Interaction, 16*(1), 1-38.

Upadhyay, U., Lancashire, G., Moser, C., & Gomez-Rodriguez, M. (2021). Large-scale randomized experiments reveals that machine learning-based instruction helps people memorize more effectively. *npj Science of Learning*, *6*(1), 1-3.

Urbina Nájera, A. B., & Calleja Mora, J. D. L. (2017). Brief review of educational applications using data mining and machine learning. *Revista Electrónica de Investigación Educativa*, *19*(4), 84-96.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

Van Eck, R. (2007). Building artificially intelligent learning games. In D. Gibson, C. Aldrich, and M. Prensky, (Eds.), *Games and simulations in online learning: Research and development frameworks* (pp. 271-307). IGI Global.

VanLehn, K. (1988). Student modeling. *Foundations of intelligent tutoring systems*, *55*, 78.

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

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

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development, 83*(4), 1137-1144.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Walker, H. M. (2004). Commentary: Use of evidence-based interventions in schools: Where we've been, where we are, and where we need to go. *School Psychology Review*, *33*(3), 398-407.

Walton, G. M., Logel, C., Peach, J. M., Spencer, S. J., & Zanna, M. P. (2015). Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *Journal of Educational Psychology*, *107*(2), 468-485.

Wang, X. (2013). Modeling entrance into STEM fields of study among students beginning at community colleges and four-year institutions. *Research in Higher Education*, *54*(6), 664-692.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernandez-Lobato, J. M., ... & Zhang, C. (2021). Results and Insights from Diagnostic Questions: The NeurIPS 2020 Education Challenge. *arXiv preprint arXiv:2104.04034*.

Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., ... & Wang, S. (2020). Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 6153-6161).

Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, *35*(1), 101-120.

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

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*(1), 87-125.

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing, 34*, 16-36.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences*, *8*(4), 327-353.

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist*, *24*(4), 345-376.

Wladis, C., Conway, K. M., & Hachey, A. C. (2016). Assessing readiness for online education--Research models for identifying students at risk. *Online Learning*, *20*(3), 97-109.

Yang, S. J. H. (2021). Guest Editorial: Precision education - A new challenge for AI in education. *Educational Technology & Society, 24*(1), 105–108.

Yang, T. Y., Baker, R. S., Studer, C., Heffernan, N., & Lan, A. S. (2019). Active learning for student affect detection. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 208 - 217). International Educational Data Mining Society.

Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. *arXiv preprint arXiv:1705.08804*.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontemp, J., Yang, S. M., Carvalho, C. M., … Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*, 364-369.

Yukselturk, E., & Bulut, S. (2007). Predictors for student success in an online course. *Educational Technology and Society, 10*(2), 71−83.

Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., & Chi, M. (2019). Hierarchical reinforcement learning for pedagogical policy induction. In *International conference on artificial intelligence in education* (pp. 544-556). Springer, Cham.

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

Situating AI (and Big Data) in the Learning Sciences: Moving Toward Large-Scale Learning Sciences

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
McNamara, Danielle	Arizona State University	
Arner, Tracy	Arizona State University	
Butterfuss, Reese	Arizona State University	
Mallick, Debshila	Rice University	
Lan, Andrew	University of Massachusetts Amherst	
Roscoe, Rod	Arizona State University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) December 20, 2022

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

rtificial Intelligence in STEM Education: The Paradigmatic Shifts in Research, Education, and Technology, pp. 2

DOI or URL to published work (if available) doi.org/10.1201/9781003181187

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 Scien through **[Grant number]** R305N2100 to **Institution]** Arizona State U.The opinions expressed are those of the authors and do not represent views of the **[Office name]** Institute of Education Scie or the U.S. Department of Education.