Hershkovitz and Alexandron *Research and Practice in Technology Enhanced Learning* (2025) 20:20
https://doi.org/10.58459/rptel.2025.20020

Research and Practice in Technology Enhanced Learning

RESEARCH Free and Open Access

A review of main issues regarding validity, reliability, generalizability, transferability, and applicability of log-based measurement of computer-assisted learning

Arnon Hershkovitz [1] * and Giora Alexandron [2]

*Correspondence:
arnonhe@tauex.tau.ac.il
School of Education,
Tel Aviv University,
30 Haim Levanon St., Tel Aviv,
Israel 6997801
Full list of author information is available at the end of the article

Abstract

Log analysis has become a common methodology in the research of computer-assisted learning. Using this method, variables to measure various aspects of learning are computed from the data that is stored in computer-assisted learning environments' log files; these files document fine-grained data on student interaction with the learning system, and are updated automatically, continuously, and unobtrusively. However, besides challenges that any empirical investigation faces, log-based studies face some other, unique challenges. Despite their methodological importance, these distinctive challenges have not yet discussed in a comprehensive manner. In this review paper, we critically examine issues of validity, reliability, generalizability and transferability, and applicability of log-based analysis. We do so by covering relevant theoretical aspects, and demonstrating them via past research. We conclude with practical recommendations for researchers in the fields of Learning Analytics and Educational Data Mining.

Keywords: Log file analysis, Learning analytics, Validity, Reliability, Generalizability, Transferability, Applicability, Relevance

Introduction

Log analysis has become an acceptable methodology in the research of computer-assisted learning over the last 15 years or so, as part of the establishment of the Educational Data Mining and Learning Analytics communities (Romero & Ventura, 2020). While implementing this methodology, data is drawn from learning environment log files (aka clickstream), which can hold a fine-grained documentation of learners' interactions with the system. Such data is logged continually and unobtrusively, hence is considered by many

scholars as representing learning processes in a more authentic way than educational data that is collected in more traditional fashion, e.g., via exams, self-report questionnaires, or observations. Therefore, log-based data has served as a popular means to inform numerous research questions in countless contexts, covering various subject matters and many different populations (Hershkovitz & Alexandron, 2019).

Over the years, log-based research in education has involved the study of different aspects of student engagement with digital learning environments, including cognitive, meta-cognitive, affective, behavioral, and social aspects (Hu & Li, 2017). Such research involves variables that are measured directly from log files in a way that extends traditional measures of learning. These variables are commonly based on students' mere interaction with the learning environment, hence rely on, e.g., page views, time between actions, or number of (re-)attempts to correctly complete a task (Champaign et al., 2014).

However, this approach entails some significant challenges that may impact the reliability, validity, applicability, and generalizability of learning assessment (Alexandron et al., 2019; Baker, 2019). Previous attempts to discuss these issues were laying the ground for a meaningful discussion of such challenges (Baker, 2019; Pelánek, 2020), however we could not find a comprehensive review of the most important issues related to ensuring the quality of log-based measurements. This is the gap we aim to bridge in the current paper. Specifically, we critically review the issues of validity, reliability, transferability, generalizability, and applicability, which lie at the heart of any empirical scientific research.

Note that we use the term "computer-assisted learning", which refers to any learning setting in which computers are involved. Think, for example, on a class working on a unit within an online learning environment in the school's computer lab, with the teacher present in the room, as part of their school day schedule; now think of the same students working on the very same unit, but this time – each completes it as a homework assignment, meaning they are allowed to work on it whenever and wherever is convenient to them. Both settings are considered as computer-assisted learning. That the different contexts should be take into consideration while designing data collection and analyzing the log files of that system is one of the takeaways from this paper, and is demonstrated whenever applicable.

Importantly, our main goal is not to justify the non-use of log-based education research; on the contrary, we would like to promote the acceptance of such methods by encouraging scholars in the field to consider the quality of their research, and by supplying scholars who are not familiar with this field with the means to evaluate such studies. To meet this goal, we were broadly led by the following research questions: What are important aspects of research quality that should be emphasized when discussing log-based research of computer-assisted learning? We took a narrative review approach, specifically a critical review (Sukhera, 2022), hence examining existing empirical studies while interpreting them from a point of view of research quality, based on a framework that we would develop

here below (in the third section, Assessing Log-Based Measurement of Computer-Assisted Learning). However, the studies that are reviewed here are not to be seen as a representative sample of all empirical, log-based explorations of learning; rather, they were selected—after a thorough, extensive search—as demonstrating the points that were raise while reading many other papers.

Background

As our main goal is to discuss the main issues of ensuring the quality of log-based research, we will first review the body of literature on the topic. First, it is relevant to present the issue of research quality assessment, specifically in the context of log-based studies; then, we highlight the process of calculating learning-related variables based on data drawn from log files.

Assessing quality of research

Research quality has been defined in many ways. Some of these refer to research design, e.g., whether it was a "true" experiment that incorporated randomized controlled trial, and to the way findings-related statements are actually supported by data and are visible, substantiate, and acceptable (Akkerman et al., 2008; Wortman, 1994). Going beyond assessing research design and presentation of findings, we would like to assess the quality of the very measures used in log-based educational research, and following – the quality of findings that are based on such measures. Therefore, we follow previous approaches to assess methodological aspects in the social sciences, e.g., qualitative studies or psychophysiological techniques, and will focus on four main criteria: validity, reliability, generalizability and transferability, and applicability (Ali & Yusof, 2011; Leung, 2015; Wang & Minor, 2008).

In his keynote at the 9th International Learning Analytics & Knowledge Conference (Tempe, Arizona, USA), Ryan Baker, one of the founding parents of the Educational Data Mining community, presented a few challenges/problems for the future of educational data mining (Baker, 2019). Baker referred to challenges related to transferability of student models from one system to another, effectiveness of using learning analytics in practice, interpretability of prediction models, applicability of knowledge tracing models to real-life, and generalizability of learning analytics models across learning systems and across populations. Baker clearly identified some pitfalls in learning analytics, doing so in the context of prediction models of learning-related behavior in intelligent tutoring systems. Of course, challenges also exist in a broader sense, as Pelánek (2020) suggested; his point of view was rather more high-level, presenting three main problems with learning analytics, related to embedded trade-offs in many learning environments (e.g., between mastery and under-practice, or between engagement and learning), methodological issues (e.g., the choice of evaluation metrics, or inherent biases when designing a research), and scalability

(i.e., scaling-up learning environments and learning analytics solutions, considering the wider world population). We continue these important discussions, aiming at presenting a comprehensive framework for quality assessment of log-based educational research.

From log files to learning assessment

Over the last three decades or so, research of computer-assisted learning has used various types of data, e.g., student information systems, sensors—like physiological signals wristbands, eye trackers, or electroencephalogram (EEG) caps—and questionnaires. However, the most common approach to study computer-assisted learning—specifically when doing so on a large-scale—has been to use data collected through log files of digital learning environments (Elmoazen et al., 2023; Kew & Tasir, 2022; Samuelsen et al., 2019). Today, this methodology is at the epicenter of a few international education research communities, like Educational Data Mining, Learning Analytics, Artificial Intelligence in Education, and Learning@Scale. Research in this field covers digital learning settings of various scales, from single digital environments used by a few students to campus-wide learning management systems and world-wide MOOCs (Massive Online Open Courses) used by millions of learners.

Log files of digital learning environments—also commonly referred to as activity logs, clickstream, or trace data—document each learner's action in three dimensions: The action taker (who?), the action itself (what?), and the action time (when?) (Ben-Zadok et al., 2009). These data have served the basis for calculating numerous variables referring to cognitive, meta-cognitive, affective, and social aspects of learning (Hershkovitz & Alexandron, 2019). As any process of measuring latent variables, log-based measurement is prone to errors; more than that, the common practice in this field of measuring learning-related variables by the construction of predictive models increases the complexity of avoiding errors (Bergner, 2017). Indeed, it has been suggested that findings from log-based studies are not easily replicated to other datasets or across contexts (Andres et al., 2004; Farrow et al., 2019), and that learning analytics results may be severely biased due to hidden sub-populations, cheating, pre-knowledge, and more (Alexandron et al., 2019; Baker & Hawn, 2022). Our main goal in this paper is to point out to possible flaws in the process of measuring learning-related behavior based on data from log files, for raising awareness to this issue among researchers, readers and practitioners.

Assessing log-based measurement of computer-assisted learning

In this section, we will present the main aspects of research quality. For each aspect, we will first explain it in a broader context, then will discuss it in the context of log-based measurement of computer-assisted learning; while doing so, we will give relevant examples from past research where required test was carried out or violated. These aspects and their purposes are summarized in Table 1.

Table 1 Summary of the research quality aspects discussed in this paper

Aspect	Purpose
Validity	
Face Validity	The extent to which the instrument is expected to provide reasonable outcomes
Content Validity	Whether the tool covers the various dimensions of the construct being measured
Criterion-Related Validity	The extent to which a measure obtained by the tested tool agrees with other measures it is expected to agree with
Construct Validity	How well are empirical evidence using the tool aligned with the theory of the construct it is intended to measure
Learner Identity	Are results based on real learner data?
Reliability	Consistency of a measuring instrument
Generalizability	The extent to which findings apply to the whole population from which the sample was drawn
Transferability	The extent to which findings apply in new contexts
Applicability	The extent to which findings are relevant to situations in the real world

Validity

Validity of a measurement can be defined as the ability to measure what is claimed to be measured (Dick & Hagerty, 1971). It is also often referred to as truthfulness, and may be conceptualized as measuring an attribute that exists and with variations in the attribute causally produce variation in the measurement outcomes (Borsboom et al., 2004). Therefore, a valid measure allows to draw conclusions from the data without logical errors (Vaughn & Daniel, 2012). It is commonly agreed that validity should be looked at from four different points of view: 1) Face validity, which refers to the extent to which the instrument is expected to provide reasonable outcomes; 2) Content validity, which aims to test whether the tool covers the various dimensions of the construct being measured; 3) Criterion-related validity, that is, the extent to which a measure obtained by the tested tool agrees with other measures it is expected to agree with; and 4) Construct validity, which tests how well are empirical evidence using the tool aligned with the theory of the construct it is intended to measure (Cohen et al. 1996). When discussing validity of log-based learning-related behaviors, we should consider all these aspects, by testing both the operationalization of variables—that is, the definition and mechanism that drives their actual computation from digital traces—and how the operationalized measure behave.

Face validity

Face validity, in the context of our discussion, may refer to the examination of the mere operationalization of variables, that is, to the way they are being calculated from the logged traces. This should be evaluated and justified explicitly, in order to make sure that the calculation mechanism is reasonable. For example, a recent publication in JLA reports on

a study of learner engagement with videos available in a MOOC (Zhang et al., 2022); one of the variables, aiming at measuring learners' watching behavior, was operationalized following the definition of Guo et al. (2014): Each video-watching session starts when a student hits the "Play Video" button, and ends when either of the following occurs: the student triggered an event that is not relevant to the video; the student ends the current logging session; or the video finished playing. Of course, such a measurement does not necessarily measure engagement with the video de-facto. This operationalization may be reasonable, and one can consider the reliance on a previously used operationalization enough for face validity, however this issue should be stressed out explicitly. That study is, of course, not unique in operationalizing engagement in such a way, as actions and timestamps are the bread and butter of log-based studies, and are easy-to-use for measuring various learning-related types of engagement (e.g., Deng & Benckendorff, 2017; Moubayed et al., 2020; Seidel, 2017). Importantly, predicting student engagement, based on log traces from learning management system, may be course-dependent, as different courses require different types of online engagement (Motz et al., 2019), so justifying a given operationalization a priori also has to do with the specific setting in matter.

Content validity

Testing for content validity means that the measurement covers all the aspects it should (and only those). This is mostly relevant when attempting to measure a high-level, complex constructs, like engagement, creativity, or self-regulated learning (SRL), just to name a few. Creativity, for example, consists of different dimensions (depending on the theoretical framework relied on); in Hershkovitz, Sitman, et al. (2019), the framework for measuring creativity required regarding to four dimensions: fluency, flexibility, originality, and elaboration; however, measuring it included only one dimension (originality).

Engagement is another complex construct that is challenging to operationalize and measure from log files (Alexandron et al., 2023). As was noted in that paper, which examined the impact of the assessment design on student engagement, "We note that most theoretical models of engagement would view this as a narrow measurement of engagement. For example, within the framework of Fredricks et al. (2004) that defined "three categories of engagement – behavioural, cognitive, and emotional – it mainly falls into the 'Behavioural' one […]" (p. 147).

In the case of SRL—a core conceptual framework to understand the complex, multi-faceted process of learning—there is no agreement on a single model, however there is a consensus among researchers that this process has multiple phases (Panadero, 2017). SRL has been studied and promoted using LA (Araka et al., 2020), however a literature review of LA studies on SRL revealed that most studies only focused on a limited set of components of this construct (Viberg et al., 2020).

Relevant to this is the very choice of indicators and their combination when building an LA-based model—i.e., a priori to their operationalization—and the idea that these should be well justified. A justification of an optimal combination of indicators usually occurs when examining an output model. This stage becomes challenging when using advanced Artificial Intelligence-based algorithms, e.g., Deep Learning, where explainability level is low. To that end, Explainable Artificial Intelligence (XAI) has been emerged as a promising approach, and specifically Explainable Learning Analytics (XLA) (De Laet et al., 2020; Khosravi et al., 2022). Taking these approaches, one uses various methods to make it clear which of the indicators was found to be prominent in the model, and what is its impact on the model, which can help in testing for context validity.

To conclude, when measuring multi-dimensional constructs via log files, it should be explicitly stated which dimensions should be included, which are measured in practice, and if there is a discrepancy between these two sets – why is it so.

Criterion-related validity

While testing for criterion-related validity, it is expected that the log-based measure of a construct will be checked for its associations with other measures with which it should agree, including of other measures of the same construct. This was done, for example, regarding engagement, which was measured by both log file indicators and self-report measures; it was found that the strength of associations between these two types of measurement may depend on the timing of the log-based measurement (Tempelaar et al., 2020). Misalignment between log-based and self-report measures was also reported regarding other constructs, e.g., achievement goal orientation or self-regulated learning (Choi et al., 2023; Salehian Kia et al., 2021); examined in the context of testing for construct validation, these findings may be interpreted as a violation of this test.

Construct validity

Construct validity testing is focused on the extent to which the tested measuring tool agrees with relevant theoretical foundations of that construct. Such agreement can be demonstrated by, e.g., an expected distribution of the behavior measured, or a changing pattern along repeated measures. D'Mello and Graesser's work on the dynamics of affective states during learning can be considered as involving testing for construct validity, as they assumed an a-priori, theory-driven model of the changes in students' affective states; as their empirical, log-based evidence agreed with that hypothesized model (to an acceptable degree), it can be seen as a validation of the measuring tool itself (D'Mello & Graesser, 2012).

Learner identity

Another core issue that is related to validity of log file-based measurement, is learner identity. Even when convinced that the measures that we use indeed measure what we think they do, we may still face a deeper problem: Are learners really who we think they are? In the case of MOOCs, Alexandron et al. (2017) pointed out to the relatively common phenomenon of learners who operate multiple accounts for collecting correct answers, which are then used by the learners' main account for credit. This – and other types of cheating, which is much more common than educators tend to believe (McCabe et al., 2010) – can certainly harm the validity of any learning-related measure. Indeed, in a later study, Alexandron et al. (2019) replicated a highly-cited learning analytics paper, and demonstrated that the high achievements that the replicated paper ascribed to 'active learning' could actually be explained by cheating. More generally, this research underlines the sensitivity of log-based educational research to a (typically implicit) crucial assumption that such research tends to make: that student achievements can be explained by their observed behavior, as captured by the log files.

Relevant to the issue of authentic learner identities is the use of simulated data. Due to the difficulty of obtaining large data sets, it has been a common practice to use simulated students that behave according to pre-defined models or rules, and are utilized as agents that simulate the interaction of human learners, leaving digital traces that are stored in the log files (Desmarais et al., 2010; Hershkovitz et al., 2017; Pelánek, 2017). The use of simulated students is often seen as "essential, since the implementation and test of new features in adaptive educational systems require substantial amounts of financial, human resources and time" (Dorça, 2015, p. 320). To put it simply, in such cases, the research population is made of imaginary students. This method has been implemented to test, research, and develop various modules of pedagogy, assessment, personalization, and adaptation (Käser & Alexandron, 2024; McCalla & Champaign, 2013). However, the validity of findings from studies implementing this methodology is debatable; as Cristea and Okamoto (2001) clearly put it, "it is questionable if a simulated student, built entirely according to the predicted student model, will not simply always generate good results, which, translated into real life, might result in completely different real evolutions of the human students" (p. 415). Thus, it is not surprising that Badiee and Kaufman (2015) reported that most of the student teachers who used simSchool – an online classroom simulation which is based on simulated students – considered it unrealistic. A recent systematic literature review, covering 176 published papers that used simulated learners (2010-2019), found that a large part of the reviewed studies did not validate the simulation in any way (Käser & Alexandron, 2024). With the emerging popularity of Generative AI models as a means of generating educational data, it is expected that the use of artificial students for studying the behavior of real ones will become more prevalent.

Reliability

Reliability is the overall consistency of a measuring instrument, that is, the degree to which it produces similar outcomes under different occasions of measurement, different persons who are involved with the instrument or with its scoring, or different editions of the instrument (Dick & Hagerty, 1971; Livingston, 2018). To give a simple example, if we would like to measure a 4th-grade student's knowledge of addition of fractions, using a pen-and-paper test, we would expect to get somewhat similar outcomes when administering this test at 9am or at 4pm (assuming no meaningful intervention for these topics was given in between).

When it comes to measuring computer-assisted learning, which is frequently flexible in its timing and location, reliability testing should explicitly consider these factors. Think of a MOOC which you can learn anytime, from anywhere, hence one's log files may document actions that took place in different times and different locations. However, timing and location of learning may impact the very way by which learners interact with digital learning environments, and therefore are important for consideration. Timing of learning has been found as an important factor that impacts learning-behavior. For example, learners' pace of activity may be affected by the hour in which learning happens (day or night), or by the timing within a learning session (beginning or end) (e.g., Hershkovitz & Nachmias, 2009), which may directly impact learning acquisition. Space was also found to potentially impact learning, as students who learn at home tend to spend more time on learning, they learn at a slower pace and score higher than students who learn in school (Ben-Zadok et al., 2010). This means that a single model for measuring a given construct may not be sufficiently reliable to account for timing and location of learning.

Moreover, digital learning environments are often characterized, more than traditional learning materials, with graphics elements, and it has been demonstrated how graphical user interface may impact time on task or learning gains (Ben-Haim et al., 2019; Hershkovitz, Tzayada, et al., 2019). So, measures of engagement and achievements within a digital learning environment are prone to reliability violation if they do not take into consideration such design factors that may differ between tasks.

Last, analysis of fine-grained clickstream data typically includes many technical decisions that may have a significant impact on the results (Pelánek et al., 2016), and are typically hidden or transparent from the end users (Feldman-Maggor et al., 2021). For example, studies that measure variables such as "time on video" need to make decisions on when to stop interpreting idle time (time without interaction) as "watching" (Champaign et al., 2014). Thirty minutes? Fifteen? Ten? Such decisions may significantly impact the results, and may be highly sensitive to the context (e.g., length of the videos and their genre), but are often taken based on general heuristics. This too may be a reliability threat.

Generalizability and transferability

Generalizability and transferability are similar concepts, and should be carefully distinguished. Generalizability refers to the extent to which a certain finding, which is based on a certain sample, applies to the whole population from which the sample was drawn. Transferability refers to the extent to which theoretical frameworks or effectiveness of interventions apply in new contexts, hence is closely associated with replicability (de Leeuw et al., 2022; Finfgeld-Connett, 2010; The National Science Foundation, 2018). For example, think of the development of a prediction model for success in an Introduction to Computer Science U.S. college-level course—based on data drawn from a sample of students from twenty U.S. colleges—which results with a very high prediction performance. A question of generalization regarding this model could be whether it well predicts performance in this course for all U.S. college population, while a question of transferability could be whether it also applies to Advanced Computer Science college-level U.S. courses, or to Introduction Computer Science college-level course in other countries. It is often stated that generalizability is applied by researchers, who take various measures to make sure their findings are indeed generalizable, while transferability is applied by readers of research, who are invited to make connections between elements of a study and their own experience (del Cerro Santamaría, 2015); this distinction helps us to understand why (Burchett et al., 2011; Mathrani et al., 2021) transferability plays a key role in teachers' use of research in their own settings (Joram et al., 2020), and why decision-makers are often more concerned with transferability of research findings than with its validity (Burchett et al., 2011).

Generalizability of log-based research in education could be increased by considering issues related to sampling, as well as by using training and testing subgroups for performance evaluation. Regarding transferability, there could be two major obstacles. The first is that a given behavior is manifested differently in different populations. For example, Rodrigo et al. (2013) showed, relying on observations on students while using a digital learning environment, that frequencies of off-task and gaming-the-system—two behaviors that have been extensively studied over the last 15 years—are different when measured in the USA and in the Philippines (while controlling for student demographics and for the learning setting). Conversely, in another case of transferability testing, Bayesian Knowledge Tracing and carelessness were found to transfer well across demographic sections (Zambrano et al., 2024). That is, different constructs may or may not be demographics-dependent, therefore the issue of transferability should be considered when wishing to implement models or findings from one demographics to another.

Furthermore, even within the US population, it was shown—while using log-based automatic detectors—that these two behaviors were manifested differently in urban, sub-urban, and rural populations (e.g., Baker & Gowda, 2010). More broadly, between-

countries differences were observed for online learning styles (Shih et al., 2013). These differences may directly impact transferability of research findings as variables' distribution may affect their relationships with other variables.

The second obstacle relates to the way different learners engage with digital learning environments—and, generally, with digital devices—which have direct impact on log-based measurements. Indeed, international studies show the large variance that exists in ICT access, use and skills in today's world population (Dodel, 2020; Fau & Moreau, 2018; Gómez-Galán et al., 2020; Hu et al., 2018), and within nations, socioeconomic variables still explain a high proportion of the variance in digital skills, in addition to demographics and other personality traits (Dodel, 2020; Hidalgo et al., 2020; van Laar et al., 2020). As recent research shows, ICT skills are positively correlated with learning from- and with engagement in technology-enhanced learning (Bergdahl et al., 2020; Dodel, 2020; Schmid & Petko, 2019). Of course, there are many other factors that may affect the ways in which students learn online (Kauffman, 2015). Hence, the very mechanisms that drive log-based assessment of learning-related behavior may prove problematic when being transferred to contexts other than those in which they were originally defined.

In addition to these obstacles, it is essential to test for transferability based on subject matter, curricular topic, and characteristics of the digital learning environment. For example, Israel-Fishelson and Hershkovitz (2019) tested for correlations between persistence and achievements—in a large scale, log-based study (N=~26,000)—within the same cohort of learners who used a single learning environment for early programming which covers various topics; they found meaningful differences in such correlations across different topics, which demonstrates the challenges in transferring research findings from one sub-topic to another. Course design may also serve as obstacle for transferability. In a large-scale, log-based study of 158 MOOCs with 2.8M enrollments from 120 countries, Gershon et al. (2021) showed that course design has an impact on its "completion bias", that is, the reduced likelihood of a subgroup of learners defined by a certain characteristic to complete the MOOC successfully; in the case of their study, subgroups of learners were defined by language (native-English speakers of non-native-English speakers) and country (developed or developing).

Of course, different learning environments may a-priori require different types of learner engagement, which reduces the set of common variables that could be measured across them. For example, the measurement of persistence mentioned above (Israel-Fishelson & Hershkovitz, 2019) was dependent, besides on learner-related factors, on the very mechanism that enabled this persistence in the first place, i.e., a "retry" button presented once a task was completed successfully. The way this button is presented to the learner, e.g., to what extent it is easily noticeable, may impact learners' actions, hence may impact the very measurement of persistence. That is, the findings are possibly sensitive to the logic

of the learning environment and to its user-interface, and therefore their transferability should be questioned (Baker, 2007; Baker & Gowda, 2010).

Lastly, the question of transferability of log-based education research should also mention traditional teaching and learning settings, when not using educational software at all. That is, are findings from log-based education research relevant to traditional, non-computerized classrooms? For example, are students in brick-and-mortar classrooms in urban schools go off-task significantly more than students in brick-and-mortar classrooms in rural and suburban schools, as was found regarding educational software and was discussed above? (Baker & Gowda, 2010). Or, to take another example that we presented earlier, are the relationships between persistence and achievement dependent on the sub-topic studied, like was found in a large-scale log-based study? (Israel-Fishelson & Hershkovitz, 2019). Of course, we should not a-priori expect such a transfer; however, when it exists – it sheds an important light on the log-based studies, and enables extending education research using yet another methodological approach. Just as Andres et al. (2004) tested findings across different settings of computer-assisted learning, findings from log-based studies should be tested in traditional learning settings.

Applicability

Applicability is the extent to which research findings are relevant to situations in the real world. Often considered as "relevance", this issue is of importance for practitioners and policymakers, as it addresses the question of whether findings from a study could be of any help in real-world scenarios (Burchett et al., 2011; Murad et al., 2018). Some common obstacles for applicability of research findings are high cost of measurement or intervention, or high sensitivity to experimental settings (Wang & Minor, 2008).

In our case, it is important to examine findings from log-based studies in the context of both real-world use of educational software and of traditional classroom learning. Regarding educational software, it is often the case that log-based studies are being conducted on historical data without implementing the log-based models as an integral part of the studied software (e.g., Cohen et al., 2021); various commercial considerations of the companies developing this software may prevent such implementation. This situation makes it difficult to test such models in real-time settings throughout the learning process using various research designs. Therefore, applicability of such findings is questionable.

But even if such models are incorporated in educational software, various obstacles may prevent from using such software in brick-and-mortar school settings on a large scale and in a way that would make real impact. Among these obstacles are technological infrastructure, curricular and pedagogical considerations, efforts needed for content development, and digital skills of teachers and students (Hershkovitz & Alexandron, 2019). To make things more complicated, even when computer-assisted learning is being used as

part of traditional learning settings, teachers may not trust log-based measures (Nazaretsky et al., 2022), and the way by which log-based information is presented to them may impact their acceptance of it (Tenório et al., 2021) – which may potentially hinder applicability of log-based research.

General discussion

The Australian sociologist Raewyn Connell suggested that research is "simply collecting information and thinking systematically about it" (Connell, 1975, p. 1). However, over the last decades to follow Connell's romantic conceptualization of research, we have witnessed an exponential growth in number of published research. The chemist and educator Kenneth Pitzer wondered about the growth in academic research since the end of World War II, and identified three main factors for that phenomenon: societal growth that allowed less able people into research, increase in published literature, and tendency towards over-specialization (Pitzer, 1967). As recent bibliometric studies demonstrate, the rapid advancement in educational technology over the last two decades or so has been accompanied by a rapid growth in research in the fields of education and technology (Jiménez et al., 2019; Song & Wang, 2020; Wahid et al., 2020). Therefore, it is now important, more than ever, to assess the quality of research in this field.

Validity, reliability, generalizability, transferability, and applicability are the cornerstone of any research. In this paper, we examined these aspects regarding log-based research of computer-assisted learning. The use of log files is now considered as a common practice in educational research, for example under the wider umbrella of Learning Analytics. By the definition adopted by SoLAR (Society for Learning Analytics Research), Learning Analytics is the measurement, collection, analysis and reporting of data on learners and their contexts in order to understand and improve learning and their environments (Simon, 2017, p. 200). That general definition directly corresponds with the definition of Learning Sciences as "the interdisciplinary empirical investigation of learning as it exists in real-world settings, and how learning may be facilitated both with and without technology" (Packer & Maddox, 2016, p. 131). That is, log-based educational research is considered as a unique case (methodology-wise) of any study of learning, which emphasizes the need to examine it critically and carefully. Ensuring such aspects in assessing one's research is indeed considered crucial in promoting high quality education research (Evans et al., 2020).

Logged data is considered to be objective in the sense that it uninterruptedly documents learner actions; however, there is still a gap between these logged actions and the latent nature of learning that may or may not be associated with them. Much emphasis has been put on improving research evidence in Learning Analytics and its associated fields—e.g., Educational Data Mining, Artificial Intelligence in Education, and Learning@Scale—of which log-based research is an integral part (Ferguson & Clow, 2017) Following our

review, we believe that at the same time, emphasis should be placed on how such evidence is obtained. Availability of digital learning environments and their logged data is often a double-edged sword. On the one hand, it can dramatically improve the study of certain behaviors and phenomena—often in large scale—even allow for exploring some that were previously impossible to be researched; on the other hand, it may bias the research community towards a tool-centric rather than problem-centric research (Wise et al., 2021).

Conclusions and recommendations

In this paper, we critically examined the validity, reliability, generalizability, transferability, and applicability of log-based research results. These are foundational dimensions that help testing research quality. As Ferguson and Clow (2017) pointed out, there is a lack of explicit evidence for such dimensions in research quality in publication in the learning analytics, hence we see this paper as a first step towards a fruitful discussion of these important issues. Of course, we highlighted potential pitfalls in various log-based education studies, but this should not be taken as an argument against this methodology, not even against the studies we brought as examples. To the contrary, we think of this critical examination of such studies as a means to strengthen the use of learning analytics and educational data mining. As ones who have been conducting log-based studies for almost two decades, we have surely committed all these 'crimes' that we are now preaching against, therefore we are currently hoping to make good use of our hard-earned experience. Based on our analysis and our experience, we would like to suggest a few practical recommendations for improving log-based measurements of computer-assisted learning.

Base your operationalization on solid theoretical grounds

Variables extracted from log files should be based on solid theoretical frameworks, just like variables used in any other educational research. This has at least two perspectives, namely, the choice of a theoretical framework and the way the measurement is consistent with this framework (Wise & Shaffer, 2015). Following the choice of a theoretical frameworks, one must make sure that the indicators used for log-based measurement are consistent with these frameworks, and that they are comprehensive with regard to it. Explicitly justifying these choices will assist reviewers and readers to evaluate face- and content validity, and will overall help in understanding the links between theory and measurement, therefore strengthening the importance of the research.

A-priori choose multiple ways of testing your model

Whether a log-based variable or model is a good fit for the construct they aim to measure should be tested from at least two points of view. First, it should be tested for the behavior of what is measured across different contexts; second, it should be tested for associations

between what is measured and other related constructs. Decisions on these tests—in accordance with the requirements of criterion-related- and construct validity—should better be made a-priori, based on theoretical grounds. This will decrease potential biases and the use of "cherry-picking" approaches.

Replicate in different settings (not just online)

Replication is one of the strongest methods to test research findings (The National Science Foundation, 2018). It is recommended to also replicate studies in other contexts than the ones for which they were originally found. These include, among other, different populations (age-wise and culture-wise), different topics or subject matters, different learning environments, and different graphical user interface. If synthetic student data is used, it is recommended to replicate the study with authentic data. Also, we recommend replicating log-based studies in non-log-based settings. These will help increase the reliability of the measures, and the generalizability and transferability of the research findings.

Set-up a standard for educational data exchange

We recommend that relevant research communities (e.g., in the fields of Educational Data Mining, Learning Analytics, and Artificial Intelligence in Education) will agree upon standards for data exchange, which will allow for secondary analysis of already-collected data. Setting up standards will help in expanding educational data repositories (e.g., those included in the LearnSphere project, http://learnsphere.org) (Koedinger et al., 2019). Standardizing educational data may also promote research-based software development, from which students, educators, and researchers will benefit, therefore increasing applicability of log-based studies (Del Blanco et al., 2013).

Raise awareness to the need in testing for validity, reliability, generalizability, transferability, and applicability

Raising awareness to the need to assess research quality vis-à-vis the aspects reviewed here may help is make these aspects visible to- and thought of by many more scholars. This community-wise discussion should better begin with every scholar reflecting on their own work, self-debating about how to assess their research using these criteria.

Abbreviations

EEG: electroencephalogram; LA: Learning Analytics; MOOC: Massive Open Online Course; SoLAR: Society for Learning Analytics Research; SRL: Self-regulated learning; XAI – Explainable Artificial Intelligence; XLA – Explainable Learning Analytics.

Authors' contributions
Conceptualization, Arnon Hershkovitz; methodology, N/A; formal analysis, N/A; data curation, Arnon Hershkovitz and Giora Alexandron; writing—original draft preparation, Arnon Hershkovitz; writing—review and editing, Giora Alexandron. Both authors read and approved the final manuscript.

Authors' information
Prof. Arnon Hershkovitz is an Associate Professor at Department of Mathematics, Science, and Technology Education, School of Education, Faculty of Humanities, Tel Aviv University, Israel. Dr. Giora Alexandron is a Senior Scientist (Assistant Professor) at the Department of Science Teaching, Weizmann Institute of Science, Israel.

Funding
This study included no funding.

Availability of data and materials
Not applicable

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Tel Aviv University, Israel

[2] Weizmann Institute of Science, Israel

Received: 10 January 2024 Accepted: 9 July 2024
Published online: 1 January 2025 (Online First: 30 July 2024)

References

Akkerman, S., Admiraal, W., Brekelmans, M., & Oost, H. (2008). Auditing quality of research in social sciences. *Quality and Quantity*, *42*(2), 257–274. https://doi.org/10.1007/s11135-006-9044-4

Alexandron, G., Wiltrout, M. E., Berg, A., Gershon, S. K., & Ruipérez-Valiente, J. A. (2023). The effects of assessment design on academic dishonesty, learner engagement, and certification rates in MOOCs. *Journal of Computer Assisted Learning*, *39*(1), 141–153. https://doi.org/10.1111/jcal.12733

Alexandron, G., Yoo, L. Y., Ruipérez-Valiente, J. A., Lee, S., & Pritchard, D. E. (2019). Are MOOC learning analytics results trustworthy? With fake learners, they might not be! *International Journal of Artificial Intelligence in Education*, *29*(4), 484–506. https://doi.org/10.1007/s40593-019-00183-1

Ali, A. Md., & Yusof, H. (2011). Quality in qualitative studies: The case of validity, reliability and generalizability. *Issues in Social and Environmental Accounting*, *5*(1), 25–64. https://doi.org/10.22164/ISEA.V5I1.59

Andres, J. M. L., Baker, R. S., Siemens, G., Gašević, D., & Spann, C. A. (2004). Replicating 21 findings on student success in online learning. *Technology, Instruction, Cognition, and Learning*, *10*(4), 313–333.

Araka, E., Maina, E., Gitonga, R., & Oboko, R. (2020). Research trends in measurement and intervention tools for self-regulated learning for e-learning environments—systematic review (2008–2018). *Research and Practice in Technology Enhanced Learning*, *15*, 6. https://doi.org/10.1186/s41039-020-00129-5

Badiee, F., & Kaufman, D. (2015). Design evaluation of a simulation for teacher education. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015592454

Baker, R. S. (2019). Challenges for the future of educational data mining: The Baker Learning Analytics Prizes. *Journal of Educational Data Mining*, *11*(1), 1–17. https://doi.org/10.5281/ZENODO.3554745

Baker, R. S. J. D. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In M. B. Rosson & D. Gilmore (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '07* (pp. 1059–1068). ACM. https://doi.org/10.1145/1240624.1240785

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Baker, R. S. J. d., & Gowda, S. M. (2010). An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools. In R. S. J. d. Baker, A. Merceron & P. I. Pavlik Jr. (Eds.), *Proceedings of the Third International Conference on Educational Data Mining* (pp. 11–20). The International Educational Data Mining Society.

Ben-Haim, E., Cohen, A., & Tabach, M. (2019). Types of graphic interface design and their role in learning via mathematical applets at the elementary school. *Eleventh Congress of the European Society for Research in Mathematics Education*.

Ben-Zadok, G., Leiba, M., & Nachmias, R. (2010). Comparison of online learning behaviors in school vs. at home in terms of age and gender based on log file analysis. *Interdisciplinary Journal of E-Skills and Lifelong Learning*, *6*, 305–322. https://doi.org/10.28945/1317

Ben-Zadok, G., Mintz, R., Hershkovitz, A., & Nachmias, R. (2009). Examining online learning processes based on log files analysis: A case study. *Research, Reflections and Innovations in Integrating ICT in Education: Proceedings of the Fifth International Conference on Multimedia and ICT in Education*, *2*.

Bergdahl, N., Nouri, J., & Fors, U. (2020). Disengagement, engagement and digital skills in technology-enhanced learning. *Education and Information Technologies*, *25*(2), 957–983. https://doi.org/10.1007/s10639-019-09998-w

Bergner, Y. (2017). Measurement and its uses in learning analytics. In C. Lang, G. Siemens, A. Wise & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 35–48). The Society for Learning Analytics Research. https://doi.org/10.18608/hla17.003

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Burchett, H., Umoquit, M., & Dobrow, M. (2011). How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks: *Journal of Health Services Research & Policy*, *16*(4), 238–244. https://doi.org/10.1258/JHSRP.2011.010124

Champaign, J., Colvin, K. F., Liu, A., Fredericks, C., Seaton, D., & Pritchard, D. E. (2014). Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In M. Sahami, A. Fox, M. A. Hearst & M. T.H. Chi (Eds.), *Proceedings of the First ACM Conference on Learning at Scale* (pp. 11–20). ACM. https://doi.org/10.1145/2556325.2566250

Choi, H., Winne, P. H., Brooks, C., Li, W., & Shedden, K. (2023). Logs or self-reports? Misalignment between behavioral trace data and surveys when modeling learner achievement goal orientation. In I. Hilliger, H. Khosravi, B. Rienties & S. Dawson (Eds.), *Proceedings of the 13th International Learning Analytics and Knowledge Conference* (pp. 11–21). ACM. https://doi.org/10.1145/3576050.3576052

Cohen, A., Ezra, O., Hershkovitz, A., Tzayada, O., Tabach, M., Levy, B., Segal, A., & Gal, K. (2021). Personalizing mathematical content in educational applets repository: Human teacher versus machine-based considerations. *Educational Technology Research and Development*, *69*(3), 1505–1528. https://doi.org/10.1007/s11423-021-10002-x

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Publishing Co.

Connell, R. (1975). *How to do small surveys – A guide for students in sociology, kindred industries and allied trades (2nd edition)*. Flinders University.

Cristea, A., & Okamoto, T. (2001). Considering automatic educational validation of computerized educational systems. In *Proceedings - IEEE International Conference on Advanced Learning Technologies, ICALT 2001* (pp. 415–417). IEEE. https://doi.org/10.1109/ICALT.2001.943962

De Laet, T., Millecamp, M., Broos, T., De Croon, R., Verbert Leuven, K. K., & Duorado, R. (2020). Explainable learning analytics: Challenges and opportunities. In M. Scheffel, V. Kovanović, N. Pinkwart & K. Verbert (Eds.), *Companion Proceedings 10th International Conference on Learning Analytics & Knowledge* (pp. 500–510). Society for Learning Analytics Research.

de Leeuw, J. R., Motz, B. A., Fyfe, E. R., Carvalho, P. F., & Goldstone, R. L. (2022). Generalizability, transferability, and the practice-to-practice gap. *Behavioral and Brain Sciences*, *45*, e11. https://doi.org/10.1017/S0140525X21000406

Del Blanco, A., Serrano, A., Freire, M., Martinez-Ortiz, I., & Fernandez-Manjon, B. (2013). E-Learning standards and learning analytics. Can data collection be improved by using standard data models? In *IEEE Global Engineering Education Conference, EDUCON* (pp. 1255–1261). IEEE. https://doi.org/10.1109/EDUCON.2013.6530268

del Cerro Santamaría, G. (2015). Transdisciplinary technological futures: An ethnographic research dialogue between social scientists and engineers. *Technology in Society*, *40*, 53–63. https://doi.org/10.1016/j.techsoc.2014.10.005

Deng, R., & Benckendorff, P. (2017). A contemporary review of research methods adopted to understand students' and instructors' use of massive open online courses (MOOCs). *International Journal of Information and Education Technology*, *7*(8), 601–607. https://doi.org/10.18178/ijiet.2017.7.8.939

Desmarais, M. C., Pelczer, I., & Montréal, P. (2010). On the faithfulness of simulated student performance data. In R. S. J. de Baker, A. Merceron & P. I. Pavlik (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 21–30). The International Educational Data Mining Society.

Dick, W., & Hagerty, N. (1971). *Topics in measurement: Reliability and validity*. McGraw-Hill Book Company.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Dodel, M. (2020). Socioeconomic inequalities and digital skills. In D. A. Rohlinger & S. Sobieraj (Eds.), *The Oxford handbook of digital media sociology* (pp. 548–566). Oxford University Press. https://doi.org/10.1093/OXFORDHB/9780197510636.013.30

Dorça, F. (2015). Implementation and use of simulated students for test and validation of new adaptive educational systems: A practical insight. *International Journal of Artificial Intelligence in Education*, *25*(3), 319–345. https://doi.org/10.1007/s40593-015-0037-0

Elmoazen, R., Saqr, M., Khalil, M., & Wasson, B. (2023). Learning analytics in virtual laboratories: A systematic literature review of empirical research. *Smart Learning Environments*, *10*, 23. https://doi.org/10.1186/s40561-023-00244-y

Evans, C., Howson, C. K., Forsythe, A., & Corony Edwards, &. (2020). What constitutes high quality higher education pedagogical research? *Assessment & Evaluation in Higher Education*, *46*(4), 525–546. https://doi.org/10.1080/02602938.2020.1790500

Farrow, E., Moore, J., & Gašević, D. (2019). Analysing discussion forum data: A replication study avoiding data contamination. In S. Hsiao, J. Cunningham, K. McCarthy, G. Lynch, C. Brooks, R. Ferguson & U. Hoppe (Eds.), *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 170–179). ACM. https://doi.org/10.1145/3303772.3303779

Fau, S., & Moreau, Y. (2018). *Building tomorrow's digital skills: What conclusions can we draw from international comparative indicators?* UNESCO. https://www.voced.edu.au/content/ngv:80941

Feldman-Maggor, Y., Barhoom, S., Blonder, R., & Tuvi-Arad, I. (2021). Behind the scenes of educational data mining. *Education and Information Technologies*, *26*(2), 1455–1470. https://doi.org/10.1007/s10639-020-10309-x

Ferguson, R., & Clow, D. (2017). Where is the evidence? A call to action for learning analytics. In A. Wise, P. H. Winne, G. Lynch, X. Ochoa, I. Molenaar, S. Dawson & M. Hatala (Eds.), *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 56–65). ACM. https://doi.org/10.1145/3027385.3027396

Finfgeld-Connett, D. (2010). Generalizability and transferability of meta-synthesis research findings. *Journal of Advanced Nursing*, *66*(2), 246–254. https://doi.org/10.1111/j.1365-2648.2009.05250.x

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Gershon, S. K., Ruipérez-Valiente, J. A., & Alexandron, G. (2021). Defining and measuring completion and assessment biases with respect to English language and development status: Not all MOOCs are equal. *International Journal of Educational Technology in Higher Education*, *18*(1), 41. https://doi.org/10.1186/s41239-021-00275-w

Gómez-Galán, J., Vergara, D., Ordóñez-Olmedo, E., & Veytia-Bucheli, M. G. (2020). Time of use and patterns of internet consumption in university students: A comparative study between Spanish-speaking countries. *Sustainability*, *12*(12), 5087. https://doi.org/10.3390/SU12125087

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement. In M. Sahami, A. Fox, M. A. Hearst & M. T. H. Chi (Eds.), *Proceedings of the First ACM Conference on Learning @ Scale* (pp. 41–50). ACM. https://doi.org/10.1145/2556325.2566239

Hershkovitz, A., & Alexandron, G. (2019). Understanding the potential and challenges of big data in schools and education. *Tendencias Pedagógicas*, *35*, 7. https://doi.org/10.15366/tp2020.35.002

Hershkovitz, A., & Nachmias, R. (2009). Consistency of students' pace in online learning. *Educational Data Mining*, 71–80. International Working Group on Educational Data Mining. http://www.educationaldatamining.org/EDM2009/uploads/proceedings/hershkovitz.pdf

Hershkovitz, A., Knight, S., Jovanovic, J., Dawson, S., & Gasevic, D. (2017). Research with simulated data. *Journal of Learning Analytics*, *4*(1), 1–2. https://doi.org/10.18608/jla.2017.41.1

Hershkovitz, A., Sitman, R., Israel-Fishelson, R., Eguíluz, A., Garaizar, P., & Guenaga, M. (2019). Creativity in the acquisition of computational thinking. *Interactive Learning Environments*, *27*(5–6), 628–644. https://doi.org/10.1080/10494820.2019.1610451

Hershkovitz, A., Tzayada, O., Ezra, O., Cohen, A., Tabach, M., Levy, B., Segal, A., & Gal, K. (2019). Can an algorithm prepare students for tasks without knowing what the tasks are? In *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence* (pp. 754–759). IEEE. https://doi.org/10.1109/CSCI49370.2019.00143

Hidalgo, A., Gabaly, S., Morales-Alonso, G., & Urueña, A. (2020). The digital divide in light of sustainable development: An approach through advanced machine learning techniques. *Technological Forecasting and Social Change*, *150*, 119754. https://doi.org/10.1016/J.TECHFORE.2019.119754

Hu, M., & Li, H. (2017). Student engagement in online learning: A review. In F. L. Wang, O. Au, K. K. Ng, J. Shang & R. Kwan (Eds.), *Proceedings - 2017 International Symposium on Educational Technology* (pp. 39–43). IEEE. https://doi.org/10.1109/ISET.2017.17

Hu, X., Gong, Y., Lai, C., & Leung, F. K. S. (2018). The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis. *Computers & Education*, *125*, 1–13. https://doi.org/10.1016/J.COMPEDU.2018.05.021

Israel-Fishelson, R., & Hershkovitz, A. (2019). Persistence and achievement in acquiring computational thinking concepts: A large-scale log-based analysis. In S. Carliner (Ed.), *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 1002–1012). Association for the Advancement of Computing in Education. https://www.learntechlib.org/primary/p/211181/

Jiménez, C. R., Prieto, M. S., & García, S. A. (2019). Technology and higher education: A bibliometric analysis. *Education Sciences*, *9*(3), 169. https://doi.org/10.3390/EDUCSCI9030169

Joram, E., Gabriele, A. J., & Walton, K. (2020). What influences teachers' "buy-in" of research? Teachers' beliefs about the applicability of educational research to their practice. *Teaching and Teacher Education*, *88*, 102980. https://doi.org/10.1016/j.tate.2019.102980

Käser, T., & Alexandron, G. (2024). Simulated learners in educational technology: A systematic literature review and a Turing-like test. *International Journal of Artificial Intelligence in Education*, *34*, 545–585. https://doi.org/10.1007/s40593-023-00337-2

Kauffman, H. (2015). A review of predictive factors of student success in and satisfaction with online learning. *Research in Learning Technology*, *23*. https://doi.org/10.3402/rlt.v23.26507

Kew, S. N., & Tasir, Z. (2022). Learning analytics in online learning environment: A systematic review on the focuses and the types of student-related analytics data. *Technology, Knowledge and Learning*, *27*(2), 405–427. https://doi.org/10.1007/s10758-021-09541-2

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

Koedinger, K., Stamper, J., Carvalho, P., Pavlik, P., & Eglington, L. (2019). Sharing and reusing data and analytic methods with LearnSphere. In C. F. Lynch, A. Merceron, M. Desmarais & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 773–774). The International Educational Data Mining Society.

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. *Journal of Family Medicine and Primary Care*, *4*(3), 324–327. https://doi.org/10.4103/2249-4863.161306

Livingston, S. A. (2018). *Test reliability - Basic concepts*. Educational Testing Service.

Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open*, *2*, 100060. https://doi.org/10.1016/J.CAEO.2021.100060

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2010). Cheating in academic institutions: A decade of research. *Ethics & Behavior*, *11*(3), 219–232. https://doi.org/10.1207/S15327019EB1103_2

McCalla, G., & Champaign, J. (2013). Simulated learners. *IEEE Intelligent Systems*, *28*(4), 67–71. https://doi.org/10.1109/MIS.2013.116

Motz, B., Quick, J., Schroeder, N., Zook, J., & Gunkel, M. (2019). The validity and utility of activity logs as a measure of student engagement. In S. Hsiao, J. Cunningham, K. McCarthy, G. Lynch, C. Brooks, R. Ferguson & U. Hoppe (Eds.), *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 300–309). ACM. https://doi.org/10.1145/3303772.3303789

Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2020). Student engagement level in an e-learning environment: Clustering using k-means. *American Journal of Distance Education*, *34*(2), 137–156. https://doi.org/10.1080/08923647.2020.1696140

Murad, M. H., Katabi, A., Benkhadra, R., & Montori, V. M. (2018). External validity, generalisability, applicability and directness: A brief primer. *BMJ Evidence-Based Medicine*, *23*(1), 17–19. https://doi.org/10.1136/EBMED-2017-110800

Nazaretsky, T., Cukurova, M., & Alexandron, G. (2022). An instrument for measuring teachers' trust in AI-based educational technology. In A. F. Wise, R. Martinez-Maldonado & I. Hilliger (Eds.), *Proceedings of the 12th International Learning Analytics and Knowledge Conference* (pp. 56–66). ACM. https://doi.org/10.1145/3506860.3506866

Packer, M. J., & Maddox, C. (2016). Mapping the territory of the learning sciences. In M. A. Evans, M. J. Packer & R. K. Sawyer (Eds.), *Reflections on the learning sciences* (pp. 126–154). Cambridge University Press. https://doi.org/10.1017/cbo9781107707221.007

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00422

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*(3–5), 313–350. https://doi.org/10.1007/s11257-017-9193-2

Pelánek, R. (2020). Learning analytics challenges: Trade-offs, methodology, scalability. In C. Rensing, H. Drachsler, V. Kovanović, N. Pinkwart, M. Scheffel & K. Verbert (Eds.), *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 554–558). ACM. https://doi.org/10.1145/3375462.3375463

Pelánek, R., Rihák, J., & Papoušek, J. (2016). Impact of data collection on interpretation and evaluation of student models. In D. Gašević, G. Lynch, S. Dawson, H. Drachsler & C. P. Rosé (Eds.), *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 40–47). ACM. https://doi.org/10.1145/2883851.2883868

Pitzer, K. S. (1967). How much research? *Science*, *157*, 779–781. https://www.science.org

Rodrigo, M. M. T., Baker, R. S. J. D., & Rossi, L. (2013). Student off-task behavior in computer-based learning in the Philippines: Comparison to prior research in the USA. *Teachers College Record: The Voice of Scholarship in Education*, *115*(10), 1–27. https://doi.org/10.1177/016146811311501007

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/WIDM.1355

Salehian Kia, F., Hatala, M., Baker, R. S., & Teasley, S. D. (2021). Measuring students' self-regulatory phases in LMS with behavior and real-time self report. In M. Scheffel, N. Dowell, S. Joksimovic & G. Siemens (Eds.), *Proceedings of the 11th International Learning Analytics and Knowledge Conference* (pp. 259–268). ACM. https://doi.org/10.1145/3448139.3448164

Samuelsen, J., Chen, W., & Wasson, B. (2019). Integrating multiple data sources for learning analytics — Review of literature. *Research and Practice in Technology Enhanced Learning*, *14*, 11. https://doi.org/10.1186/s41039-019-0105-4

Schmid, R., & Petko, D. (2019). Does the use of educational technology in personalized learning environments correlate with self-reported digital skills and beliefs of secondary-school students? *Computers & Education*, *136*, 75–86. https://doi.org/10.1016/J.COMPEDU.2019.03.006

Seidel, N. (2017). Analytics on video-based learning. A literature review. In C. Ullrich & M. Wessner (Eds.), *Proceedings of DeLFI and GMW Workshops 2017*. CEUR-WS.org.

Shih, Y.-C. D., Liu, Y.-C., & Sanchez, C. (2013). Online learning style preferences: An analysis on Taiwanese and USA learners. *Turkish Online Journal of Educational Technology*, *12*(4), 140–152.

Simon, J. (2017). A priori knowledge in learning analytics. In A. Peña-Ayala (Ed.), *Learning analytics: Fundaments, applications, and trends* (Vol. 94, pp. 199–227). Springer, Cham. https://doi.org/10.1007/978-3-319-52977-6_7

Song, P., & Wang, X. (2020). A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years. *Asia Pacific Education Review*, *21*(3), 473–486. https://doi.org/10.1007/s12564-020-09640-2

Sukhera, J. (2022). Narrative reviews: Flexible, rigorous, and practical. *Journal of Graduate Medical Education*, *14*(4), 414–417. https://doi.org/10.4300/JGME-D-22-00480.1

Tempelaar, D., Nguyen, Q., & Rienties, B. (2020). Learning analytics and the measurement of learning engagement. In D. Ifenthaler & D. Gibson (Eds.), *Adoption of data analytics in higher education learning and teaching. advances in analytics for learning and teaching* (pp. 159–176). Springer, Cham. https://doi.org/10.1007/978-3-030-47392-1_9

Tenório, K., Lemos, B., Nascimento, P., Santos, R., Machado, A., Dermeval, D., Paiva, R., & Isotani, S. (2021). Learning and gamification dashboards: A mixed-method study with teachers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12677 LNCS* (pp. 406–417). Springer, Cham. https://doi.org/10.1007/978-3-030-80421-3_45

The National Science Foundation. (2018). *Companion guidelines on replication & reproducibility in education research: A supplement to the common guidelines for education research and development*. The National Science Foundation.

van Laar, E., van Deursen, A. J. A. M., van Dijk, J. A. G. M., & de Haan, J. (2020). Determinants of 21st-century skills and 21st-century digital skills for workers: A systematic literature review. *SAGE Open*, *10*(1), 1–14. https://doi.org/10.1177/2158244019900176

Vaughn, B. K., & Daniel, S. R. (2012). Conceptualizing validity. In G. Tenenbaum, R. C. Eklund & A. Kamata (Eds.), *Measurement in sport and exercise psychology* (pp. 33–39). Human Kinetics.

Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online learning environments. In C. Rensing, H. Drachsler, V. Kovanović, N. Pinkwart, M. Scheffel & K. Verbert (Eds.), *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 524–533). ACM. https://doi.org/10.1145/3375462.3375483

Wahid, R., Ahmi, A., & Alam, A. S. A. F. (2020). Growth and collaboration in massive open online courses: A bibliometric analysis. *International Review of Research in Open and Distance Learning*, *21*(4), 292–322. https://doi.org/10.19173/IRRODL.V21I4.4693

Wang, Y. J., & Minor, M. S. (2008). Validity, reliability, and applicability of psychophysiological techniques in marketing research. *Psychology and Marketing*, *25*(2), 197–232. https://doi.org/10.1002/mar.20206

Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, *2*(2), 5–13. https://doi.org/10.18608/JLA.2015.22.2

Wise, A. F., Knight, S., & Ochoa, X. (2021). What lakes learning analytics research matter. *Journal of Learning Analytics*, *8*(3), 1–9. https://doi.org/10.18608/jla.2021.7647

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 97–110). Russel Sage Foundation.

Zambrano, A. F., Zhang, J., & Baker, R. S. (2024). Investigating algorithmic bias on Bayesian Knowledge Tracing and carelessness detectors. In B. Flanagan, B. Wasson & D. Gašević (Eds.), *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (pp. 349–359). ACM. https://doi.org/10.1145/3636555.3636890

Zhang, J., Huang, Y., & Gao, M. (2022). Video features, engagement, and patterns of collective attention allocation. *Journal of Learning Analytics*, *9*(1), 32–52. https://doi.org/10.18608/jla.2022.7421

Publisher's Note