Review article

doi: 10.1111/jcal.12172

Journal of Computer Assisted Learning

Review of computer-based assessment for learning in elementary and secondary education

V.J. Shute & S. Rahimi

Florida State University, Florida, USA

Abstract In this paper, we review computer-based assessment for learning (CBAfL), in elementary and secondary education, as a viable way to merge instruction and assessment of students' developing proficiencies. We begin by contextualizing our topic relative to summative and formative assessment before presenting the current literature, which we categorized into the following: (a) supplementary use in classrooms, (b) web-based, and (c) data-driven, continuous CBAfL. Examples of research studies per category are provided. Findings show that using CBAfL in the classroom, via the Internet, or embedded in a game, generally enhances learning and other outcomes across a range of content areas (e.g. biology, math, and programming). One conclusion is that feedback, to be most beneficial to learning, should not be overly complex and must be used to be effective. Findings also showed that the quality of the assessment (i.e. validity, reliability, and efficiency) is unimpaired by the inclusion of feedback. The possibilities created by advances in the learning sciences, measurement, and technology have paved the way toward new assessment approaches that will support personalized learning and that can accurately measure and support complex competencies. The next steps involve evaluating the new assessments regarding their psychometric properties and support of learning.

Keywords Assessment, learning, assessment for learning, computer-based assessment for learning, elementary education, secondary education.

> Assessment should not merely be done to students; rather, it should also be done for students, to guide and enhance their learning (National Council of Teachers of Mathematics, 2000).

Assessment refers to not only systematically collecting and analysing information about a learner – what people normally think of when they hear the term assessment – but also to interpreting and acting on information about learners' understanding and/or performance in relation to educational goals (Bennett, 2011; Pellegrino, Chudowsk & Glaser, 2001), and while there is a variety of assessment types, the choice and use of an assessment should depend on the educational purpose. For example, schools make heavy use of standards-based summative assessment (also known as assessment *of* learning), which is useful for accountability purposes (e.g. unidimensional assessment for grading and promotion purposes) but only marginally useful for supporting personal learning. In contrast, learner-centred measurement models rely mostly on formative assessment, also known as assessment *for* learning, which can be very useful in guiding instruction and supporting individual learning, but may not be particularly consistent or valid. That is, a current downside of the assessment-for-learning model is that it is often implemented in a non-standardized and hence less rigorous manner than summative assessment and thus can hamper the validity and consistency of the assessment tools and data (Shute

Accepted: 12 November 2016

Correspondence: Valerie J. Shute, Florida State University, Tallahassee, Florida FL 32306-4453, USA. Email: vshute@fsu.edu

& Zapata-Rivera, 2010). This is not to say such assessments do not have value. Rather, it is a call for research to come up with new techniques and research studies to determine assessments' value and/or utility (e.g. employing a meta-analysis approach with formative assessment studies to provide an aggregate picture that cannot be seen clearly through individual cases). Strong formative assessment research is needed given changes in (a) the types of learning we are valuing today (and in the near future), as well as (b) the new, broader set of contexts in which learning is taking place. According to Bennett (2011), previous meta-analyses and the associated claims with regard to assessment for learning are flawed. Specifically, he criticizes the vague claims of effectiveness and argues for assessing effect sizes through more domain-specific approaches.

This paper reviews the literature related to a particular form of assessment – computer-based assessment for learning (CBAfL). Our focus is on its use and effectiveness in elementary and secondary education (i.e. Kindergarten through grade 12). To justify the need for such a paper, we first examined nine recent and relevant literature reviews. The reviews covered topics related to (a) assessment for learning in the classroom (Bennett, 2011; Birenbaum et al., 2015; Heitink, van der Kleij, Veldkamp, Schildkamp & Kippers, 2016), (b) summative and formative assessment in the digital age (Oldfield, Broadfoot, Sutherland & Timmis, 2012; Timmis, Broadfoot, Sutherland & Oldfield, 2015), (c) effectiveness of feedback in computer-based assessment environments (van der Kleij, Feskens & Eggen, 2015; van der Kleij, Timmers & Eggen, 2011) and (d) psychometric analysis of the performance data of simulation-based assessments (de Klerk, Veldkamp & Eggen, 2015).

Collectively, the reviews tended to focus on both assessment of and assessment for learning rather than only on assessment for learning. In addition, the reviews covered a range of education levels rather than just elementary and secondary education. For the reviews that explicitly focused on assessment for learning, some tended to be very specific (e.g. the effectiveness of feedback in a computer-based assessment environment, or the effectiveness of audience response systems). Finally, among the reviews that examined assessment for learning in general, we were unable to find any review that related specifically to CBAfL. Consequently, we believe that the current review is needed, focusing on computer-based assessment for learning *CBAfL in elementary and secondary education.*

Before reviewing the relevant literature, we first present a brief description of summative and formative assessment to set the stage – noting that we focus on the latter.

Summative and formative assessment

The two most familiar assessments are summative and formative. Summative assessment (or assessment of learning) involves using assessment information for high-stakes and/or cumulative purposes, such as for grades, promotion and certification. It is usually administered after some major event, like the end of the school year or marking period; or before a big event, like college entry. Benefits of standards-based, criterion-referenced summative assessment are that (a) it allows for comparing learner performances across diverse populations on clearly defined educational objectives and standards; (b) it provides reliable data (e.g. scores) that can be used for accountability purposes at various levels (e.g. classroom, school, district, state and national) and for various stakeholders (e.g. learners, teachers and administrators); and (c) it can inform educational policy (e.g. curriculum or funding decisions).

Formative assessment (or assessment for learning) is intended to support teaching and learning. It is incorporated directly into the classroom curriculum and uses results from learners' activities as the basis on which to adjust instruction to promote learning in a timely manner. A simple example would be a teacher giving a pop quiz to his students on some key topic, immediately analysing their scores and then refocusing his lesson to straighten out a prevalent misconception shared by the majority of students in the class. Formative assessments are usually administered more frequently than summative assessment and have shown potential for supporting learning in different content areas and for diverse audiences (e.g. Black & Wiliam, 1998; Hindo, Rose & Gomez, 2004; Schwartz, Bransford & Sears, 2005; Shute, Hansen & Almond, 2008). In addition to providing teachers with evidence about how their class is learning so that they can revise instruction appropriately, formative assessment typically provides feedback directly to the learners to help them gain insight about how to improve, and by suggesting (or implementing) instructional adjustments based on assessment results. For instance, it may allow fast learners to move on to advanced topics, and slow learners spend more time on the topic with which they are struggling (Reigeluth & Karnopp, 2013).

In general, formative assessment provides helpful information to teachers about their students' learning progress and to students so that they know where they are, where they are going and how to get there in terms of their knowledge and skills (Black & Wiliam, 2009; Brown, 2004; Lin & Dwyer, 2006). Research suggests that formative assessment results in greater achievement for most students compared with standard pedagogical approaches, especially low achievers (Black & Wiliam, 2009; Lin & Dwyer, 2006; Stiggins, 2002). Thus, formative assessment has a potentially important role to play in education.

In this paper, we use the term 'assessment for learning' rather than formative assessment because the latter is too vague and the former is currently more prevalent among researchers in the area (Bennett, 2011; Cech, 2007). Moreover, our review focuses particularly on CBAfL – as a viable way for teachers to unify assessment and instruction within a diagnostic, computer-based formative assessment system (Black & Wiliam, 1998; Shute, Leighton, Jang & Chu, 2016a).

The aim of this paper is to present findings from a literature review of CBAfL to gain a better understanding of the features, functions, interactions and links to learning, in elementary and secondary schools. The findings can also inform future work in this area.

Method

Procedure

Seminal articles in the assessment for learning literature were identified and then collected. The bibliography compiled from this initial set of research studies spawned a new collection-review cycle, garnering even more articles, and continuing iteratively throughout the review process. The following online databases were employed in this search–collection effort:

- *ERIC*: This database contains educational reports, evaluations and research from the Educational Resources Information Centre, consisting of Resources in Education Index, and Current Index to Journals in Education.
- *PsycINFO:* This database is from the American Psychological Association, which carries citations and summaries of scholarly journal articles, book chapters, books and dissertations, in psychology, education and related disciplines.
- *Web of Science*: This database includes Science Citation Index, Social Science Citation Index, Arts and Humanities Citation Index and the Conference Proceedings Citation Index in Science and Social Science.
- *Google Scholar*: This site provides a simple way to broadly search for relevant literature. One can search across many disciplines and sources (articles, theses, books and abstracts) from academic publishers, professional societies, online repositories, universities and other websites.

In addition to searching these databases, we looked at the reference sections of the relevant review papers we collected and other empirical studies to find more articles that we did not find in our database searching process.

Inclusion criteria

The focus of the search was to access full-text documents using various search terms or keywords, such as the following: CBAfL, computer-based assessment, computer-based testing, computer-assisted testing, formative assessment, assessment for learning, K-12, elementary education, secondary education and diagnostic assessment. The search was not limited to a particular date range, although slight preference was given to more recent research. We concentrated on full-text, peer-reviewed, high-quality journal articles (including review papers), dissertations, book chapters and 'other' (e.g. research reports). An initial screening of our search results yielded about 140–160 studies, then after a subsequent screening, we ended up with nine review papers and eight empirical studies as examples of CBAfL that met our inclusion criteria (i.e. relevancy to the topic, situated in the context of elementary and secondary education, ranging in geographical location, and having a sound design and methodology).

Computer-based assessment for learning

The advent of computer-based assessment for learning systems

Computers in the 1960s and early 1970s were not very powerful (e.g. black and white text-based interfaces with 8K RAM). However, a few visionary educators did see the potential for using computers as learning and assessment tools in the 1960s (e.g. Green, 1964). The first generation of computer-based learning environments consisted of a computerized version of programmed instruction, defined as any systematic and structured teaching approach which aims to reinforce some desired behaviours (Gagné & Brown, 1961). These computerized teaching systems were called computer-assisted instruction (or computer-based training; Dyke & Newton, 1972; Shute & Psotka, 1996). Starting in the 1970s, computerized testing systems were developed and studied in classroom environments. This refers to the use of computers to create and administer assessments that can be used during the school year as a supplementary learning tool (Cartwright & Derevensky, 1975). Computers generated different sets of questions using question banks and provided immediate feedback to the students (Charman & Elmes, 1998).

Such computer-based tests were used by various teachers and researchers from the 1970s through the 1990s to measure fairly simple declarative knowledge based on students' correct or incorrect responses to questions (e.g. Cartwright & Derevensky, 1975; Charman & Elmes, 1998; Mooney, 1998; Zakrzewski & Bull, 1998). For example, Cartwright and Derevensky (1975) conducted a study where they used computer-based multiple-choice quizzes to individually assess students' knowledge of the subject matter over time, with an immediate feedback mechanism in place. Results typically showed that students using the computerized tests demonstrated significantly more favourable attitudes towards computer-based instruction compared with those who used paper-and-pencil tests. Also, students perceived the computerized tests as a useful mastery learning tool.

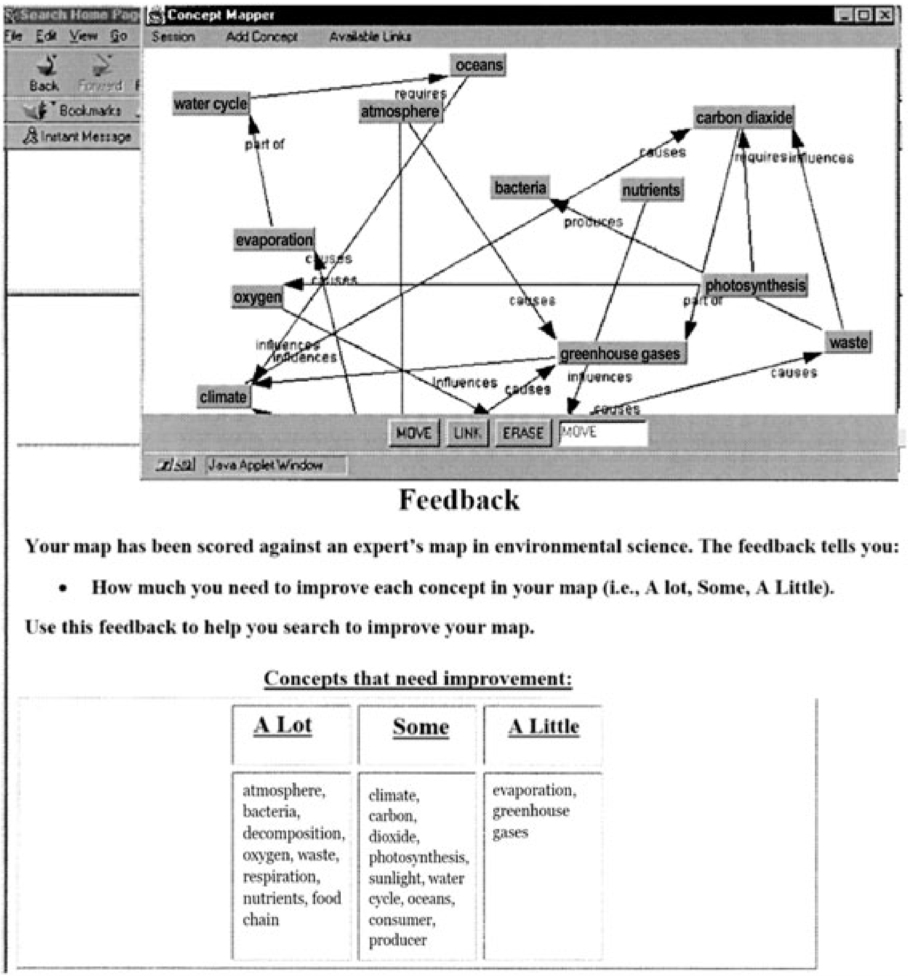
In the late 1990s, researchers began using computers to assess more complex cognitive skills like problem-solving (Baker & Mayer, 1999; O'Neil, 1999; Schacter, Herl, Chung, Dennis & O'Neil, 1999). Previously, problem-solving was measured by looking at the learner's solutions, which is an inadequate assessment of problem solving abilities (Baker & Mayer, 1999). Computer environments, with their data collection and computational capabilities, made it possible for researchers to assess complex skills more effectively than they had been able to before. For example, Schacter et al. (1999) used an Internet-based program that included four computational tools to assess problem-solving processes and performance: (a) Java Mapper – a tool for creating concept maps, (b) a Web-based learning environment where students could search for learning materials, (c) a bookmarking application that allowed students to send items they found to their knowledge map and (d) outcome feedback, which provided feedback to students based on a comparison between students' maps and an expert map. This system provided feedback to students with information about concepts that needed much improvement (i.e. no matches with the expert map), some improvement (at least 20% matches with the expert map) and little improvement (quite similar to the expert map). Figure 1 shows an example of the feedback students received.

In general, CBAfL systems in the past – from the early 1960s to the late 1990s – were used to provide computer-assisted instruction or to serve as a supplementary tool for instructional support in classrooms. As technology advanced, CBAfL systems evolved beyond just computerizing instruction or tests – for example, measuring more complex competencies like problem-solving skills (Baker & Mayer, 1999; O'Neil, 1999; Schacter et al., 1999). This trend continues today. Next, we review the current state of CBAfL in elementary and secondary education.

Today's computer-based assessment for learning systems

In the past couple of decades, advances in the learning sciences and technology have influenced new thinking and practices related to assessment for learning. For instance, advances in the learning sciences indicate that acquiring and demonstrating new knowledge and skills occurs within an environment or pedagogical context, which includes (a) learners with specific cognitive and emotional profiles and (b) tools to promote and evaluate student learning (Pellegrino et al., 2001).

In general, learning sciences have evolved from behaviourism to cognitivism and then to constructivism and situated learning (Driscoll, 2005). New pedagogical approaches have been introduced (e.g. problem-based learning, project-based learning, collaborative learning and game-based learning) that demonstrate a shift from a teacher-centred to a more learner-centred approach. In contrast with the older views of learning (i.e. learning only happens within an individual's mind), the new approaches support learning that occurs in various contexts and among the minds of different people



Feedback

Your map has been scored against an expert's map in environmental science. The feedback tells you:

- **How much you need to improve each concept in your map (i.e., A lot, Some, A Little).**

Use this feedback to help you search to improve your map.

Concepts that need improvement:

A Lot	Some	A Little
atmosphere, bacteria, decomposition, oxygen, waste, respiration, nutrients, food chain	climate, carbon, dioxide, photosynthesis, sunlight, water cycle, oceans, consumer, producer	evaporation, greenhouse gases

Figure 1 Outcome Feedback, from Schacter et al. (1999)

(Lave & Wenger, 1991). In addition, research and practices within the learning sciences are currently moving towards personalized learning (Martinez, 2002; Shute et al., 2016a) and away from the traditional one-size-fits-all model of teaching and learning. That is, no student should have to move on if they have not solidly learned a topic, and no student should be held back from learning new concepts and skills if the rest of the class is not there yet (Reigeluth & Karnopp, 2013).

Technology has also dramatically changed the environments and processes by which students learn and communicate, teachers instruct and assessments are designed and administered. Over the past several decades, there has been an explosion of new technologies (e.g. smartphones, tablets, high-speed computers, high-speed Internet, wearable devices and virtual and augmented reality) and associated research (Shute et al., 2016a) that are finding their way into education. In particular, CBAfL has directly benefited from new technologies, advances in the learning sciences, as well as measurement techniques. Paper-and-pencil tests are slowly becoming a thing of the past as assessments are now increasingly being designed as adaptive and delivered online (e.g. computer adaptive testing), employing dynamic and interactive tasks and simulations (e.g. Luecht, 2013). Items for large-scale tests are increasingly created and assembled automatically by sophisticated computer algorithms that cannot only produce items in more cost-effective ways but also enough of them to address security concerns (Shute et al., 2016a). These innovations are beginning to influence the science of assessment, allowing greater ecological validity and feedback to students related to the breadth and depth of knowledge and skills learned *in-situ*, including complex skills (e.g.

critical thinking, creativity, collaboration and problem-solving). That is, advances in technologies and their integration with assessment systems have allowed for the assessment of multidimensional learner characteristics (e.g. cognitive, metacognitive and affective; Pellegrino et al., 2001; Raymond & Usherwood, 2013) using authentic digital tasks, such as games and simulations (e.g. Azevedo, Johnson, Chauncey & Burkett, 2010; Blanchard, Wiseman, Naismith & Lajoie, 2012; Shute & Ventura, 2013).

Some technological advances that have affected education the most include the Internet, Smart boards in classrooms, online games, social networks and mobile devices that are examples of technology-rich environments. Advances in technology and the increasing application of technology-rich environments in educational settings have provoked discussions of assessment in the digital age (McFarlane, 2003) and technology-enhanced assessment (i.e. use of technologies to enhance formal or informal assessment; Timmis et al., 2015). Previously, measuring complex skills was difficult given the lack of clear and established definitions, theoretical multidimensionality of the constructs and subjectivity with regard to scoring (Shute & Wang, 2016). Simple multiple-choice questions or self-report measures are not well suited for assessing such complex skills. Instead, innovative assessments are needed to accurately measure and support these hard-to-measure constructs.

After screening and evaluating an initial collection of CBAfL studies, we selected eight research studies (from 2008 to 2016), in three non-mutually-exclusive categories for our review (Table 1): (a) supplementary use of CBAfL in class, (b) Web-based CBAfL and (c) data-driven and continuous CBAfL.

The studies cover different types of CBAfL, from different countries, across various subject areas and using different outcome measures.

Supplementary use of computer-based assessment for learning in class

Conducting frequent formative assessments and providing feedback for a large group of students can be very time-consuming for teachers, thus making it unappealing in practice (Burns, Klingbeil & Ysseldyke, 2010; McGuire, 2005). This is where computers can be quite useful. CBAfL helps by providing informative feedback to students and making their learning experience more personalized within both face-to-face and online classes

Table 1. Studies Included in the Review of Current Computer-based Assessments for Learning.

	Context				Unit of analysis		
First author and Year	Country	Education setting	Subject	CBAfL tool	Schools	Students	Focus
			Supplementary use of CBAfL in class				
Maier, Wolf, and Randler (2016)	Germany	SE	Multiple subjects	Moodle	—	216	Feedback type
Shute (2008)	USA	SE	Algebra	ACED	—	268	Adaptive CBAfL/feedback types effectiveness
Rodrigues and Oliveira (2014)	Portugal	SE	History	AssiStudy	—	723	Free-text assessment
Burns (2010)	USA	EE	Math	Accelerate Math	360	—	effect of CBAfL on learning
			Web-based CBAfL				
Wang (2011)	Taiwan	SE	Biology	PDA-WATA	—	123	Peer-driven CBAfL and self-regulated learning
Koedinger, McLaughlin, and Heffernan (2010)	USA	EE	Math	ASSISTments	—	1240	Effect of CBAfL on learning
			Data driven and continuous CBAfL				
Berland, Davis, and Smith (2015)	USA	SE	IPRO programming	AMOEBA	—	95	Improving learning via data-driven paring
Shute (2016b)	USA	SE	Problem-solving	Use your brainz game	—	55	Validation of stealth assessment

CBAfL = computer-based assessment for learning; SE = Secondary Education; EE = Elementary Education; PDA-WATA = peer-driven assessment module of the Web-based assessment and test analysis; ACED = adaptive content with evidence-based diagnosis; IPRO = I can PROgram.

(Burns et al., 2010; Salvia, Ysseldyke & Witmer, 2012; Ysseldyke & McLeod, 2007).

There are two main aims of CBAfL in classroom contexts: (a) provide appropriate and timely feedback to students and (b) personalize learning. Improving the quality of feedback and the way it is delivered are critically important components of learning (Farrell & Rushby, 2015; Shute, 2008; Timmis et al., 2015; van der Kleij, Eggen, Timmers & Veldkamp, 2012), and computers can help to accomplish these two goals (Thelwall, 2000). We now describe four examples of CBAfL used in classroom environments.

The first CBAfL system we review was developed in Moodle (a Web-based learning management system) and used in ten Biology classes (grades 6 and 7) in Germany ($n = 261$). Maier, Wolf, and Randler (2016) investigated the effects of feedback type on student achievement. The two types of feedback were as follows: (a) *elaborated* feedback, which includes explanatory/instructional information and (b) *verification* feedback, which simply confirms whether or not the learner responded correctly to a question. Students in each class were randomly assigned to one of three conditions: elaborated-feedback group (T1), verification-feedback group (T2) and no-feedback group (control). Students in groups T1 and T2 used their CBAfL system after two topics were covered in class (for 45 min on each of two separate days). The students in the control group just read text related to the biology topics they learned in the same session. The findings revealed some unexpected results. That is, the students in the T2 group receiving verification feedback demonstrated significantly higher posttest scores of their conceptual knowledge compared with the students in the T1 group receiving elaborated feedback. Moreover, there were no significant differences between either of the two treatment conditions with the control group. This was contrary to the researchers' hypothesis, as well as to the findings from other research studies examining computer-based formative feedback (e.g. Hattie & Timperley, 2007; Shute, 2008; van der Kleij et al., 2011).

Subsequently, the researchers divided the T1 group's data into two subgroups: (a) those who actually used the elaborated feedback and perceived it as helpful ($n = 46$) and (b) those who did not use the feedback and/or did not find it helpful ($n = 33$). The new findings showed that the students in the elaboration subgroup who attended to the feedback (T1A) performed the same as those in verification feedback group (T2) on the conceptual knowledge posttest, and both groups (T1A and T2) scored higher than the subgroup who received elaboration feedback but did not use it (T1B) as well as higher than the control group.

Maier et al. (2016) concluded that the elaborated feedback provided to the T1 group did not work as intended because of the way it was designed (i.e. the feedback text was too long and detailed). Moreover, in some cases, the elaborated feedback may not have been warranted, especially for some of the simpler concepts and for the more motivated students. So while feedback is an important part of any learning system (Hattie & Gan, 2011; Hattie & Timperley, 2007), it should be provided in manageable units (Shute, 2008) to ensure its use and to prevent cognitive overload, which seems to be the case in this example. Finally, as Boud and Molloy (2013) suggest, the feedback cycle should conclude with a tangible effect on students' learning. If the feedback is not actually used by students for any reason, it will not be effective.

The second CBAfL system is called adaptive content with evidence-based diagnosis (ACED; Shute et al., 2008). ACED assesses students' knowledge and skill levels of Algebra I content (i.e. arithmetic, geometric and other recursive sequences) and combines adaptive task sequencing with elaborated feedback to support student learning. The authors tested three main features of ACED: feedback type (elaborated vs. verification, like the Maier et al., 2016 study), task sequencing (adaptive vs. linear) and competency estimation. The specific research questions were as follows: (a) Is elaborated feedback (i.e. task-level feedback that provides explanations for incorrect responses) more effective for student learning than simple feedback (verification only)?, (b) Does adaptive sequencing of the assessment tasks have any impact on student learning compared with linear sequencing? and (c) Does the provision of task-level feedback affect the validity, reliability and/or efficiency of the assessment?

Shute and colleagues conducted a controlled evaluation testing 268 high-school students who were randomly assigned to one of four conditions: (a) elaborated feedback (verification and explanation) + adaptive sequencing of items ($n = 71$), (b) simple feedback (verification) + adaptive sequencing of items ($n = 75$), (c) elaborated feedback + linear sequencing ($n = 67$) and (d) the no-treatment control group ($n = 55$). All four groups completed a pretest, then the three treatment groups used ACED (tailored per group based on the treatment condition) for 1 h. Students in the control group read content that was unrelated to math for 1 h. Finally, all four groups completed the post-test for 20 min. Overall and not surprisingly, the results showed that students who used ACED (i.e. all three groups, combined) scored significantly higher on the post-test, holding pretest constant, than the control group [$F(1, 266) = 6.00$; $p < 0.02$]. The effect size was $d = 0.38$. Further analysis showed that students in group 1 (elaborated feedback + adaptivity) scored significantly higher on the post-test compared with those in group 2 (simple feedback + adaptivity) (Mean Difference = 5.74; $SE = 2.09$; $p < 0.01$) and the control group (Mean Difference = 6.36; $SE = 2.52$; $p < 0.02$). Thus, ACED helped students improve their Algebra I knowledge and skills primarily because of the receipt of elaborated feedback. This finding aligns with previous findings about elaborated feedback (e.g. Hattie & Timperley, 2007; van der Kleij et al., 2011), but contrasts with what Maier et al. (2016) found. The sequencing of tasks in ACED (adaptive vs. linear) did not show significant effects on learning. Finally, the authors showed that the quality of the assessment (i.e. validity, reliability and efficiency) was unimpaired by the provision of feedback which suggests that assessments in other settings (e.g. state-mandated tests) might be augmented to support student learning with instructional feedback without jeopardizing the primary purpose or psychometric properties of the assessment.

Our third example of a CBAfL system, called assisted study (AssiStudy), was developed and tested by Rodrigues and Oliveira (2014). It was used in a high-school history course in Portugal, but the authors note that it was developed as a generic and flexible system that may be applied to other content areas. One attractive feature of this system is that it can analyse text input from students. Given that assessing free-text responses is a very time-consuming and difficult task for teachers, AssiStudy helps teachers create and administer tests, and monitor students' progress during the course by providing different types of information about the students (e.g. scores from training exams taken by students to help teachers understand students' degree of preparedness prior to the final exam, and questions answered incorrectly by the majority of students to help teachers adjust the difficulty level of similar questions on future exams).

AssiStudy automatically generates tests for students to use as practice, receive immediate feedback and prepare for the final exam. The evaluation of students' free-text responses to history-related questions is based on semantic and syntactic similarities between the student answers and various reference/expert answers stored in the system. The system uses natural language processing techniques to make students' answers easier to process (Rodrigues & Oliveira, 2014). To make this comparison, the system pre-processes the content using natural language processing techniques, distilling student responses down to their main points. The feedback presented to the student provides explanations relative to the questions on which the student did not respond well, aiming to correct any misconceptions, bugs and other errors. Findings show that the instructors' and CBAfL system evaluations of students' free-text responses were significantly correlated ($r = 0.88$). Furthermore, the average percentage of students that passed the final exam was greater for those who used AssiStudy than for those who did not [$t(533) = 57.65$, $p < 0.05$]. This suggests that using AssiStudy has a positive impact on students' performance involving free-text responses to questions.

Finally, accelerated math (AM; Renaissance Learning, 1998) is the name of a popular CBAfL system used in math classrooms. AM was created in 1998 and is still being used today in different schools (from elementary through high school). It was designed to improve math achievement through individualized drill and practice and to help teachers provide appropriate feedback and monitor their students' progress (Burns et al., 2010). Burns et al. conducted a study to test if the schools using AM showed a higher percentage of students scoring at a proficient level than schools that did not use AM (or any other CBAfL program). Data were collected from 360 randomly selected schools across four states: Florida, Minnesota, New York and Texas. The researchers selected the respective states' summative assessments as the dependent variable: (a) Florida Comprehensive Assessment Test, (b) Minnesota Comprehensive Assessment, (c) New York State Testing Program and (d) Texas Assessment of Knowledge and Skills.

The findings showed that the schools employing the AM program ($n = 240$) showed a higher percentage of students scoring in the proficient range than the control schools ($n = 120$) that did not use this system [$F(2, 357) = 19.27$, $p < 0.001$], even after using reading test scores as a covariate. The schools that used AM for five or more years ($M = 76.22$, $SD = 13.28$) had a higher percentage of students scoring proficiently than the control group ($M = 64.03$, $SD = 17.64$; $d = 0.78$). Moreover, schools that used the AM program for 5 years or more showed a slightly higher percentage of students who scored proficiently compared with schools that used the program for 1 to 4 years ($M = 72.49$ and $SD = 15.52$; $d = 0.25$). Additional studies have been conducted using AM in math classes (e.g. Gaeddert, 2001; Lambert, Algozzine & Mc Gee, 2014; Powell, 2014; Ysseldyke, Spicuzza, Kosciolek & Boys, 2003), which have similarly shown positive effects on students' math achievement.

Among the four studies described above earlier using CBAfL in the classroom, it seems that CBAfL enhances learning across a range of content areas (biology, math and history). And although elaborated feedback is generally found to be more helpful than simple verification feedback (Hattie & Timperley, 2007; Shute, 2008; van der Kleij et al., 2011), if it is too detailed or complex, the feedback can be useless to students. Therefore, the first takeaway message from this section is that feedback, to benefit learning, should not be overly complex. Additionally, longer-term use of a CBAfL system appears to be more beneficial to students learning than shorter-term use (e.g. Burns et al., 2010). Thus, the second takeaway message is that the duration of using CBAfL systems like AM influences results – that is, the more students use such systems, the better they perform on end-of-year standardized tests (Burns et al., 2010; Gaeddert, 2001; Lambert et al., 2014; Powell, 2014; Ysseldyke et al., 2003). Next, we discuss the literature related to Web-based CBAfL.

Web-based and computer-based assessment for learning CBAfL

In general, assessment *of* learning tends to be more prevalent than assessment *for* learning, particularly in online than in face-to-face settings (e.g. Hewson, 2012; Pachler, Daly, Mor & Mellar, 2010). However, online-learning environments provide many opportunities for incorporating formative assessment as a tool to increase teacher–student and student–student interactions. As Gikandi, Morrow and Davis (2011) pointed out, students in face-to-face classes have many opportunities to interact with their peers and their teacher as they seek and receive feedback. This suggests that teachers have more opportunities to informally assess students' progress in face-to-face settings than in online settings (Gikandi et al., 2011).

Purely Web-based learning environments lack the type of immediate, face-to-face interaction occurring naturally in classroom settings, which can be detrimental to learning (Haythornthwaite, Kazmer, Robins & Shoemaker, 2000). However, Web-based assessment for learning systems can play an important role in (a) keeping students engaged with the course material and (b) providing the means for students to monitor their progress in the course. Such online assessments can also help to increase the level of meaningful interactions, which can help prevent additional learning problems related to online settings (e.g. the isolation effect, self-regulation issues, lack of motivation and retention in the course). CBAfL tools can be used solely online, or they can be blended with face-to-face instruction.

Web-based assessment for learning may seem similar to our previous category (supplementary CBAfL in the classroom), but there are some differences. For example, in the Web-based CBAfL examples discussed the following paragraphs, students access assessments via the Internet and learn about a concept (e.g. evolution and fractions) independently and at their own pace. However, when CBAfL tools are used as supplements in a classroom, students first learn the concepts and then complete a CBAfL in class, usually the same day. In short, Web-based CBAfL provides more autonomy for students even if used as a supplementary tool in class. We now examine two studies involving Web-based CBAfL systems used in elementary and secondary education.

Wang (2011) developed a Web-based assessment for learning system called the peer-driven assessment module of the Web-based assessment and test analysis (PDA-WATA). This Web-based assessment system aims to help students improve their self-regulatory learning behaviour (e.g. when learning a concept in a stand-alone e-learning module on the Internet). PDA-WATA provides five strategies to promote self-regulated learning, which is particularly relevant in Web-based e-learning environments where students need to learn without an instructor or with the instructor as a facilitator: (a) *Adding answer notes*, to explain the reason for choosing a certain option as the correct answer; (b) *Stating confidence*, to assert one's level of confidence about a particular answer and associated answer notes; (c) *Reading peer answer notes*, to see other students' answer notes; (d) *Recommending peer answer notes*, to endorse valuable answer notes for use by other students; and (e) *Querying peers' recommendations of personal answer* *notes*, to help students seek additional recommendation information from PDA-WATA about their own answer notes by other peers (i.e. the number of times a student's answer note was recommended by other peers).

Peer-driven assessment module of the Web-based assessment and test analysis's effectiveness on students' learning self-regulatory behaviour in a Web-based e-learning system was examined through a quasi-experimental design study in a junior high school in Taiwan (four seventh grade classes; $n = 123$). These four classes were randomly assigned to either the PDA-WATA group ($n = 63$) or the normal Web-Based test (N-WBT) group ($n = 60$). Students in both groups were provided two lessons on evolution (part of a Biology course) in an online e-learning environment equipped with Web-based CBAfL (i.e. the assessments were part of the e-learning lessons) for 2 weeks, comprising six sessions. Throughout the six sessions, students in both groups went through the same e-learning content and learned the lessons online, mostly independently, with the teacher only facilitating the learning process. The only difference between the two e-leaning environments was the Web-based assessments. The treatment group used PDA-WATA that supports students' self-regulated learning. Students had to respond to five questions (out of 15) that were randomly presented by PDA-WATA. If they could correctly answer a question three consecutive times, then that question would not show up again. This process continued until all 15 items were answered, and the system marked the student as having passed the formative assessment. Moreover, students could take the assessment as many times as they wanted, and they could take it anywhere and at any time. However, students could only read peers' answer notes 12 times in the study. The other group (i.e. N-WBT) used a Web-based formative test where the students would answer all questions (15 items) at a time, and at the end, they received the results with feedback (i.e. verification feedback plus explanation of the correct answer). The students in the N-WBT were also able to take the assessment as many times as they wanted until they answered all items correctly.

Wang recorded the number of times students in both groups actually used their Web-based assessment tool to examine the degree to which PDA-WATA and N-WBT motivated students, which was presumed to affect learning. The researcher also administered the learning process inventory (LPI; Gordon, Dembo & Hocevar, 2007) to all students – a 7-point Likert scale questionnaire measuring self-regulated learning behaviour. This scale includes the sub-scales of self-monitoring, deep strategy use, shallow processing, persistence and environmental structuring. The LPI was administered as both a pretest and post-test to measure students' self-regulatory behaviours before and after using the Web-based formative assessments (i.e. PDA-WATA and N-WBT). In addition to this scale, Wang developed a summative assessment targeting the evolution concepts, which were used to measure the effectiveness of the e-learning system. Students in both groups completed the LPI and the evolution assessment as a pretest, completed 2 weeks of instruction and then completed the LPI and evolution assessment as a post-test.

Results showed that the number of times students elected to use the PDA-WATA system ($M = 41.73$, $SD = 14.11$) was significantly greater than the usage patterns of the N-WBT students ($M = 9.87$, $SD = 11.50$) [$t(122) = 13.69$, $p < 0.01$]. This finding was in line with the researcher's hypothesis. In addition, the PDA-WATA students showed significantly higher LPI post-test scores (i.e. self-regulatory behaviours) compared with those in the N-WBT group [$F(1, 120) = 12.31$, $p < 0.01$]. Wang then divided student data within each group into low vs. high subgroups based on a median split of the LPI data. Comparing students with high-LPI scores in the PDA-WATA group ($M = 135.98$, $SD = 20.22$) with their high counterparts in the N-WBT group ($M = 124.09$, $SD = 24.20$) showed a medium-to-high effect size ($d = 0.53$). Similarly, but more striking, comparing students with low-LPI scores in the PDA-WATA group ($M = 98.69$, $SD = 15.74$) with those low in the N-WBT group ($M = 78.48$, $SD = 26.47$) showed a large effect size ($d = 0.93$). These results suggest that CBAfL tools like the PDA-WATA are more effective than normal CBAfL tools like N-WBT with regard to improving students' self-regulatory behaviours. Finally, the analysis of the summative assessment data on evolution content showed that students in the PDA-WATA group had significantly higher scores than students in the N-WBT group [$F(1, 120) = 19.15, p < 0.01$]. In general, it seems that a CBAfL system which encourages a student to be mindful and reflective during an assessment (e.g. articulating the rationale for making a specific response and one's confidence in the answer) improves self-regulatory skills, which in turn improves learning.

The second Web-based CBAfL system we review is called ASSISTments, developed in 2003 to deliver math assessments coupled with timely assistance in the form of instructional support to students (Heffernan & Heffernan, 2014). ASSISTments provide formative feedback to students (about their answers and overall performance), teachers (about their students' progress, strengths and weaknesses), as well as school administrators and parents in an online platform (see https://www.assistments.org). Students can use ASSISTments to improve their math knowledge, independently through drill and practice with automated immediate feedback. Moreover, ASSISTments provide descriptive statistical reports for the instructors that can help them make data-driven adjustments to their lesson plans based on students' performance (e.g. emphasize a topic on which students did not do well). ASSISTments therefore can be used as a purely online CBAfL tool by students and/or as an online supplementary tool in class by the instructors.

Koedinger, McLaughlin and Heffernan (2010) evaluated the effectiveness of using ASSISTments for 1 year on students' end-of-year test scores on the Massachusetts Comprehensive Assessment System (MCAS). Koedinger et al. used a quasi-experimental research design in this study with a sample of seventh grade students (n = 1240) from four middle schools in Massachusetts. Among the students, 985 were regular students (79%), and 255 were special education students (21%). There were three treatment schools that used ASSISTments for one school year: TA (n = 372), TB (n = 322), TC (n = 253); and one control school that used traditional textbook activities for one school year: C (n = 293). Students' sixth-grade MCAS scores were used as a pre-assessment of their incoming knowledge, and their MCAS seventh-grade scores were used as the dependent variable. Students who completed 60 or more items (each item is called an ASSISTment) were defined as high-usage students, those who completed less than 60 items were defined as low-usage students and those who did not use ASSISTments at all were defined as non-usage students. Based on the researchers' approximation, 60 items reflect about 2 h of content coverage.

The results of an ANCOVA analysis of variance (using sixth-grade MCAS scores as the covariate) with condition (i.e. three treatment schools and one control) and education group (regular vs. special education) as factors showed a significant main effect for both condition [F(1, 1235) = 12.3, $p < 0.001$] and education group [F(1, 1235) = 119.4, $p < 0.001$], as well as a significant interaction effect between condition and education group [F(1, 1235) = 6.6, p = 0.01]. The adjusted post-test means of the three treatment groups and control group differed by 3.3% (d = 0.23), a small effect size. Both regular and special education students using ASSISTments performed better than their counterparts in the control group. Moreover, the difference was significant for the special education students assigned to Assistments [F(1, 1235) = 11.44, $p < 0.001$; d = 0.50], but not for the regular students [F(1, 1235) = 1.16, p = 0.28; d = 0.08]. These results suggest that CBAfL systems like ASSISTments can improve students' learning, particularly for students in special education.

The authors additionally investigated the effect of ASSISTment usage on students' performance on the MCAS test. Results of an analysis of variance (high vs. low vs. non-usage) of regular students showed a significant main effect for student usage [F(2, 744) = 15.05, $p < 0.001$], with high-usage students (M = 61.65, SE = 0.75) performing better than the low-usage students (M = 58.06, SE = 0.52) and non-usage students (M = 54.99, SE = 0.99) on the seventh-grade MCAS exam. These results suggest that the more students use ASSISTments, the better they perform on summative assessments like the MCAS. The same usage analysis, however, with data from the special education students, was not significant.

Various studies have investigated the impact of Web-based CBAfL on students' self-regulated learning strategies (Mahroeian & Chin, 2013; Wang, 2007; Wang, 2010), knowledge acquisition and enjoyment (Tsai, 2013), and student achievement (Wang, Wang, Wang & Huang, 2006). Overall, findings show positive effects of Web-based formative assessments on different dependent variables under investigation. The two papers highlighted in the current review similarly reported positive findings for Web-based CBAfL, but more empirical research is needed, particularly to identify features and functions that are more or less effective for different types of students. Web-based environments typically do not have a synchronous face-to-face teacher, which may negatively impact student learning (Hewson, 2012). Effective CBAfL can potentially counter some of the negative effects. Finally, an advantage of Web-based CBAfL systems is their accessibility and ease of use. That is, they can be accessed via many devices (e.g. desktop computers, smartphones and tablets) both solely online by students (i.e. students can go online anywhere they want, practice and receive feedback) and blended with face-to-face instruction (i.e. instructors ask students use any CBAfL system online in class or at home). Moreover, data from Web-based CBAfL systems can be quickly accessed by instructors, students, administrators and parents (with reports personalized to each stakeholder's needs) and can be used immediately to address students' learning issues – not when it is too late. Our final section examines data-driven and continuous CBAfL – which may be employed in either face-to-face or online environments.

Data-driven and continuous computer-based assessment for learning

As mentioned earlier in this review, with the advent of new technologies (e.g. technology-rich environments) and advances in the learning and assessment sciences, we are beginning to be able to accurately assess complex skills and consequently make more informed and targeted decisions on improving student learning (Shute et al., 2016a; Timmis et al., 2015). These assessments are data-driven and can be continuous – updating in real time and accumulating across time. In this section, we review two recent studies using data-driven, continuous CBAfL, related to learning analytics (LA) and game-based assessment.

Learning analytics

Educational data mining (EDM) and LA are two similar fields of study that recently emerged in the past decade or two (Baker & Yacef, 2009). Their main purpose is to analyse large-scale data – or 'big data' – collected from learning environments and learner interactions therein (e.g. Papamitsiou & Economides, 2014; Siemens & Baker, 2012). EDM and LA researchers aim to better understand learners and the settings they learn in, and to optimize the learning processes and outcomes within those environments (Tempelaar, Heck, Cuypers, van der Kooij & van de Vrie, 2013). These data analytic fields hold promise for creating truly personalized learning through extensive and ongoing analyses of big data that the learners produce. Despite the similarities between these two fields, they have some subtle differences (e.g. in LA, leveraging human judgement using automated discovery is key, while in EDM, automated discovery using human judgment is key; Siemens & Baker, 2012).

Several of the main researchers in these two fields have conducted literature reviews on both EDM and LA (Baker, 2011; Baker & Yacef, 2009; Papamitsiou & Economides, 2014; Romero & Ventura, 2007; Siemens & Baker, 2012; Sin & Muthu, 2015; Vahdat et al., 2015). A common thread in these review papers is that both EDM and LA reflect the era of data-intensive or data-driven educational approaches that can lead to high-quality personalized and well-informed learning experiences for learners (Vahdat et al., 2015). We now review a study that used an LA tool to improve novice programmers' learning.

Berland, Davis and Smith (2015) tested an LA tool called AMOEBA. The tool was specifically developed to help secondary school teachers manage the collaboration among their students in a programming environment called I can PROgram (IPRO; a visual programming environment based on 'drag and drop' functionality to create virtual soccer player robots to play with other virtual robots). AMOEBA provides analyses of students' programming behaviours in the IPRO environment. The data from IPRO are stored in an online server. Teachers use the information from AMOEBA to set up small groups (pairs) of novice programming students to work together on programming problems in class.

The participants of this study were secondary school students (grades 7–12) across eight different classes (n = 95). All students used an iOS-based device (iPad or iPhone) to interact with the IPRO environment. Two types of data were examined in this study: *programming* data which included students' programming log files and *pairing* data which was provided by AMOEBA in a graphical representation showing students' program similarities. Similarities were evaluated by parsing each program into a tree with several sub-trees, and the sub-trees are evaluated for similarities. Students' program data were analysed to explore how pairing with AMOEBA impacted the complexity of students' code, program novelty and program quality.

AMOEBA provides recommendations for student pairings based on programming similarities (in the paper, this is referred to as being within one another's zone of proximal development). As students work together on their programs, AMOEBA updates its pairing recommendations in real time. Teachers then can either keep the pairings as they are (if the pair is functioning successfully) or change pairings, based on AMOEBA's recommendation.

In the study, students completed several phases of a programming activity: (a) familiarization with IPRO, which was an introduction to the program delivered by the teacher; (b) individual programming (pre-paired situation), in which students had to create virtual soccer player robots on their own; (c) paired programming (paired situation), where students continued their individual programming while sitting in pairs that were suggested by the AMOEBA's analyses, and helping each other if they faced any programming problems; and (d) post-paired, if the initial pair assignment was no longer similar as judged by AMOEBA. That is, while students were programming in pairs, the instructor monitored students' program similarities via AMOEBA, and a student would get re-assigned whenever stronger connections between their program and another student's program appeared on the AMOEBA's interface, or whenever the connection between the paired nodes disappeared (i.e. when the programs no longer were similar). While there were no restrictions on the number of repairings, the instructor sought to minimize the number of repairings to avoid negative, disruptive effects on student learning. This process continued for about 90 min – a full class session.

Berland et al. (2015) analysed the effectiveness of this pairing methodology using AMOEBA in the three aforementioned situations (i.e. pre-paired, paired and post-paired) on four outcome variables: *rarity* (on a continuous scale of 0 to 1 showing the uniqueness of the program compared with other programs, where 0 = common and 1 = novel); *quality* (where the soccer robot, created by a student, competes in simulation games against another robot to score points, and higher scores = higher quality of the program); *depth* (the number of levels of the parse-tree – branching conditions – in the written program to be compiled); and *specificity* (the program's length that is increased by any function added to the program, for example, 'AND', 'OR' and 'IF'). All metrics changed significantly over the 90-min class time in each of the three conditions (pre-paired, paired and post-paired) except for the quality metric. The results for the four metrics are as follows: rarity [$F(3, 91) = 3.32$, $p < 0.05$], quality [$F(3, 91) = 2.44$, $p = 0.09$], depth [$F(3, 91) = 13.16$, $p < 0.001$] and specificity [$F(3, 91) = 9.45$, $p < 0.001$]. The authors explained that the quality metric was not significant because as the pairing continued and programs became more complex, the ratio between the scores for and against the students' robots became smaller. In fact, the quality of the programs was increasing, but the *p* value of the quality metric became insignificant. The authors reported that all four metrics increased after students were paired and continued to increase after pairing. Implications of these findings regarding collaborative learning for novice programming classes can be useful for both teachers (to orchestrate student pairings in an effective way) and students (to write effective code with help from their peers). Next, we examine game-based assessment as another type of data-driven, continuous CBAfL.

Game-based assessment

Video games can be used as a vehicle for assessment. McClarty et al. (2012) have pointed out that games are inherently ongoing assessments. Similarly, Gee and Shaffer (2010) have argued that in conjunction with developing games for learning, we should focus on developing games for assessment purposes. Game-based assessment refers to a particular use of games that captures real time, in-game activities as evidence for making inferences about competencies of interest (Kim, Almond & Shute, 2016). Stealth assessment (Shute, 2011) refers to evidence-based, ongoing and unobtrusive assessments that can capture, measure and support the growth of targeted competencies. This kind of assessment can be used in games to adapt to players' performance levels with scaffolding, appropriate feedback and other types of support and also to provide appropriately challenging levels (Shute, Ke & Wang, in press). As the player interacts with the game, stealth assessment (which is embedded deeply within the game) analyses patterns of actions using the game's log file to estimate the player's competencies and make claims about them. Stealth assessment maintains a competency model per player and continuously updates it as the player interacts with the game. Information from the competency model is used to adapt the game to the player's ability level and create a personalized learning/playing experience.

To illustrate, Shute, Wang, Greiff, Zhao and Moore (2016b) developed and embedded a stealth assessment in a game called *Use Your Brainz* (a slightly modified version of the popular game *Plants vs. Zombies 2*) to measure middle-school students' problem-solving skills

($n = 55$). The researchers started by developing a competency model based on an extensive literature review of problem-solving skill. This competency model has four main facets: (a) analysing givens and constraints, (b) planning a solution pathway, (c) using tools and resources effectively and efficiently and (d) monitoring and evaluating progress. Next, they identified particular indicators (i.e. observable actions) that would provide evidence for each variable in the competency model (i.e. 32 in-game indicators: seven for analysing givens and constraints, seven for planning a solution pathway, 14 for using tools and resources effectively and efficiently, and four for monitoring and evaluating progress). Finally, they created Bayesian networks to accumulate the incoming data from game play in real time and update beliefs relative to the facets in the competency model.

In *Use Your Brainz*, players must position various plants on their lawn to prevent zombies from reaching their house. Each of the plants has different attributes. For example, some plants (offensive ones) attack zombies directly, while other plants (defensive ones) slow down zombies to give the player more time to attack the zombies. A few plants generate ‘sun’, an in-game resource needed to produce more plants. The challenge of the game comes from determining which plants to use and where to place them in order to defeat all the zombies in each level of the game. The researchers were able to incorporate stealth assessment in this game given their collaboration with Glasslab (who obtained the source code for *Plants vs. Zombies 2* and made direct changes to the game as needed – to specify and collect particular information in the log files).

In the study, students played the game across three consecutive days (about 1 h/day). To validate the stealth assessment estimates of problem-solving skill, on the fourth day, students completed two external problem-solving measures: Raven’s progressive matrices (Raven, 1941) and MicroDYN (Wüstenberg, Greiff & Funke, 2012), and additionally completed a demographic questionnaire (e.g. age, gender and gaming history). Results indicated that the problem-solving estimates derived from the game significantly correlated with the external measures: Raven’s ($r = 0.40$, $p < 0.01$) and MicroDYN ($r = 0.41$, $p < 0.01$). These correlations suggest convergent validity. Suggested next steps include running a larger validation study and developing tools to help educators interpret the results of the assessment, which will subsequently support the development of problem-solving skills at the facet level.

Based on our review, the data-driven and continuous CBAfL techniques/methods (i.e. EDM, LA and game-based assessment) are relatively new, and yet, they provide many opportunities for improving learning (e.g. effective pairing of students for collaborative learning activities and accurately measuring hard-to-measure constructs like problem-solving). With the current pace of advancements in technologies and the learning sciences, we expect – in the near future – to see high-quality digital learning environments equipped with data-driven and ongoing CBAfL systems that can seamlessly and accurately measure a wide range of complex competencies (including knowledge, skills, affective states and so on) and successfully provide effective learning supports based on students’ needs, thus delivering personalized learning experiences for all.

Summary and discussion

In this article, we reviewed research related to CBAfL in elementary and secondary education, spanning a range of content areas and outcomes. In the past, such technology was used in elementary and secondary settings to provide simple computer-assisted instruction (e.g. Dyke & Newton, 1972; Shute & Psotka, 1996), computer-generated tests using question banks with immediate feedback to the students (Charman & Elmes, 1998) and/or to serve as a supplementary tool for instructional support (Cartwright & Derevensky, 1975). The older CBAfL systems were used mainly to assess students’ simple declarative knowledge based on their responses to questions, and studies examining their effectiveness showed generally positive results (e.g. Cartwright & Derevensky, 1975; Charman & Elmes, 1998; Mooney, 1998; Zakrzewski & Bull, 1998). From the late 1990s to the present, advances in the learning sciences, technologies and measurement methods have collectively spawned new digital learning environments (e.g. technology-rich environments) in which complex competencies can be measured and supported (McFarlane, 2003; Shute et al., 2016a; Timmis et al., 2015). In this article, we reviewed the current state of CBAfL across three categories: (a) supplementary use of CBAfL in class, (b) Web-based CBAfL and (c) data-driven and continuous CBAfL.

In the first category (supplementary use of CBAfL in the classroom), the research we presented focused on similar goals as in the earlier CBAfL systems (i.e. provide timely feedback to students and personalize learning experiences). However, the quality of the feedback and the way it is delivered have been dramatically improved compared with past CBAfL systems (Farrell & Rushby, 2015; Shute, 2008; Timmis et al., 2015; van der Kleij et al., 2012). The main findings from this research include the following: (a) feedback should be designed such that students actually use it rather than ignore it (Maier et al., 2016) and delivered in manageable units (Shute, 2008), (b) well-designed, elaborated feedback is more helpful than simple verification feedback (Rodrigues & Oliveira, 2014; Shute et al., 2008) and (c) CBAfL systems can enhance learning across a range of content areas (biology, math and history).

In the second category (Web-based CBAfL), we examined research related to Web-delivered assessments for learning. Our findings indicate that generally, assessment *of* learning tends to be more prevalent than assessment *for* learning in online-learning settings (e.g. Hewson, 2012; Pachler et al., 2010). However, CBAfL tools can enhance Web-based learning through keeping students engaged with the course material, providing the means for students to monitor their progress in the course, and increasing the level of meaningful interactions. Self-regulatory skills can also be supported by encouraging students to be mindful and reflective when responding to questions in a Web-based assessment for learning system (e.g. articulating the rationale for making a specific response and specifying one's confidence in the answer) which in turn can improve learning (Wang, 2011). Moreover, there is a positive relationship between the use of Web-based CBAfL tools (e.g. ASSISTments) and students' performance on summative standardized tests (Koedinger et al., 2010; Wang, 2011).

The third category of research we examined looked at data-driven and continuous CBAfL. Research in this area is on track to accomplish high quality, ongoing, data-driven assessment for learning. The data generated by students' interactions with technology-rich environments can be analysed and aggregated using new methods (e.g. EDM and stealth assessment in games) to find hidden learning and error patterns and confirm learning progress, with the goal to enhance learning more effectively and efficiently than was possible in the past (Baker & Yacef, 2009; Papamitsiou & Economides, 2014; Siemens & Baker, 2012). The main findings in relation to this category suggest that data-driven and ongoing CBAfL (e.g. EDM, LA and stealth assessment) hold great promise for creating high-quality and personalized learning experiences for elementary and secondary learners through extensive and ongoing analyses of the data that the learners produce in various digital learning environments (Siemens & Baker, 2012; Vahdat et al., 2015). In addition, certain LA tools like AMOEBA can effectively manage students' collaborative learning activities (e.g. paring students based on the real-time data analysis) (Berland et al., 2015). Finally, complex competencies (e.g. problem-solving skill) can be accurately measured by data-driven continuous assessment techniques like stealth assessment (Shute et al., 2016b) and those diagnostic data may be used to support those skills.

Based on the trends we presented in our review, CBAfL is likely to continue to improve in the personalization of learning across a variety of contexts, such as face-to-face classrooms, online environments and informal settings like museums and after-school programs. We also expect that innovative CBAfL techniques will move beyond the laboratory and into the mainstream, and we will no longer have to rely solely on high-stake tests for assessing students' knowledge and skills (Shute et al., 2016a). The boundaries between instruction, learning and assessment will eventually become blurred. As a result, students will not have to worry about taking exams, teachers will not have to spend time preparing and grading the exams, and parents will enjoy seeing their children engaged with learning (Reigeluth & Karnopp, 2013; Shute, Rahimi & Sun, in press). Toward this end, researchers, educators and policymakers will need to embrace a model that includes the ongoing gathering and sharing of data for continuous improvement of learning and teaching (National Education Technology Plan, NETP, 2016). Additional research needs to be conducted on developing systems to deliver valid, reliable, fair and cost-effective CBAfL to accurately measure and improve complex competencies across various disciplines in the near future.

References

Azevedo, R., Johnson, A., Chauncey, A., & Burkett, C. (2010). Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In M. Khine & I. Saleh (Eds.), *New science of learning* (pp. 225–247). Amsterdam: Springer.

Baker, E., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior*, *15*(3), 269–282.

Baker, R. (2011). Data mining for education. In B. McGaw, P. Peterson & E. Baker (Eds.), *International encyclopedia of education* (3rd ed.). Oxford, UK: Elsevier.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, *1*(1), 3–17.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25.

Berland, M., Davis, D., & Smith, C. P. (2015). AMOEBA: Designing for collaboration in computer science classrooms through live learning analytics. *International Journal of Computer-Supported Collaborative Learning*, *10*(4), 425–447.

Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A., … Wyatt-Smith, C. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. *Policy Futures in Education*, *13*(1), 117–140.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (Formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5–31.

Blanchard, E. G., Wiseman, J., Naismith, L., & Lajoie, S. P. (2012). A realistic digital deteriorating patient to foster emergency decision-making skills in medical students. *In 12th IEEE International Conference on Advanced Learning Technologies,* (pp. 74–76). Rome, Italy: IEEE Communications Society. doi:10.1109/ICALT.2012.44

Boud, D., & Molloy, E. (2013). Decision-making for feedback. In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education* (pp. 202–217). London: Routledge.

Brown, S. (2004). Assessment for learning. *Learning and Teaching in Higher Education*, *1*(1), 81–89.

Burns, M. K., Klingbeil, D. A., & Ysseldyke, J. (2010). The effects of technology-enhanced formative evaluation on student performance on state accountability math tests. *Psychology in the Schools*, *47*(6), 582–591.

Cartwright, G. F., & Derevensky, J. L. (1975). An attitudinal study of computer-assisted testing as a learning method. *Psychology in the Schools*, *13*(3), 317–321.

Cech, S. J. (2007). Test industry split over “formative” assessment. *Education Week*, *28*(4), 1–15.

Charman, D., & Elmes, A. (1998). *Computer based assessment (volume 1): A guide to good practice*. University of Plymouth: SEED Publications.

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education*, *85*, 23–34.

Driscoll, M. P. (2005). *Psychology of learning for instruction* (3rd ed.). Boston: Allyn and Bacon.

Dyke, B. F. V., & Newton, J. M. (1972). Computer-assisted instruction: Performance and attitudes. *The Journal of Educational Research*, *65*(7), 292–293.

Farrell, T., & Rushby, N. (2015). Assessment and learning technologies: An overview. *British Journal of Educational Technology*, *47*(1), 160–120.

Gaeddert, T. J. (2001). Using accelerated math to enhance student achievement in high school mathematics courses. (Master’s thesis), Friends University, Seattle, Washington.

Gagné, R. M., & Brown, L. T. (1961). Some factors in the programming of conceptual learning. *Journal of Experimental Psychology*, *62*(4), 313–321.

Gee, J. P., & Shaffer, D. (2010). Looking where the light is bad: Video games and the future of assessment. *Phi Delta Kappa International EDge*, *6*(1), 3–19.

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351.

Gordon, S. C., Dembo, M. H., & Hocevar, D. (2007). Do teachers’ own learning behaviors influence their classroom goal orientation and control ideology? *Teaching and Teacher Education*, *23*(1), 36–46.

Green, B. F. (1964). Intelligence and computer simulation. *Transactions of the New York Academy of Sciences*, *27*(1 Series II), 55–63. doi:10.1111/j.2164-0947.1964.tb03486.x.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In P. Alexander & R. E. Mayer (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487.

Haythornthwaite, C., Kazmer, M. M., Robins, J., & Shoemaker, S. (2000). Community development among distance learners: Temporal and technological dimensions. *Journal of Computer-Mediated Communication*, *6*(1). doi:10.1111/j.1083-6101.2000.tb00114.x.

Heffernan, N. T. & Heffernan, C. L. (2014). The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497.

Heitink, M., van der Kleij, F., Veldkamp, B., Schildkamp, K., & Kippers, W. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review*, *17*, 50–62.

Hewson, C. (2012). Can online course-based assessment methods be fair and equitable? Relationships between students' preferences and performance within online and offline assessments. *Journal of Computer Assisted Learning*, *28*(5), 488–498.

Hindo, C., Rose, K., & Gomez, L. M. (2004). Searching for steven spielberg: Introducing iMovie to the high school english classroom: A closer look at what open-ended technology project designs can do to promote engaged learning. In *Proceedings of the 6th International Conference on Learning Sciences* (pp. 609–609). Mahwah, NJ: Erlbaum.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing*, *16*(2), 142–163.

Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*, *43*(4), 489–510.

Lambert, R., Algozzine, B., & Mc Gee, J. (2014). Effects of progress monitoring on math performance of at-risk students. *British Journal of Education, Society and Behavioural Science*, *4*(4), 527–540.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge university press.

Lin, H., & Dwyer, F. (2006). The fingertip effects of computer-based assessment in education. *TechTrends*, *50*(6), 27–31. doi:10.1007/s11528-006-7615-9.

Luecht, R. M. (2013). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59–78). New York, NY: Routledge.

Mahroeian, H., & Chin, W. M. (2013). An analysis of web-based formative assessment systems used in e-learning environment. In *13th International Conference on Advanced Learning Technologies (ICALT),* 77-81. IEEE.

Maier, U., Wolf, N., & Randler, C. (2016). Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education*, *95*, 85–98.

Martinez, M. (2002). What is personalized learning. *The E-Learning Developers' Journal: Strategies and Techniques for Designers, Developers, and Managers of eLearning,*

McClarty, K. L., Orr, A., Frey, P. M., Dolan, R. P., Vassileva, V., & McVay, A. (2012). *A literature review of gaming in education* (Technical Report). New Jersey: Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Lit_Review_of_Gaming_in_Education.pdfPearson Publishing.

McFarlane, A. (2003). Assessment for the digital age. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 261–266.

McGuire, L. (2005). Assessment using new technology. *Innovations in Education and Teaching International*, *42*(3), 265–276.

Mooney, G. (1998). Some techniques for computer-based assessment in medical education. *Medical Teacher*, *20*(6), 560–566.

National Council of Teachers of Mathematics (NCTM) (2000). *Principles and standards for school mathematics*. Reston VA: National Council of Teachers of Mathematics.

National Education Technology Plan (NETP). (2016). Future ready learning: Reimagining the role of technology in education. Retrieved from http://tech.ed.gov/files/2015/12/NETP16.pdf

Oldfield, A., Broadfoot, P., Sutherland, R., & Timmis, S. (2012). Assessment in a digital age: A research review. (Online report No. 6). Bristol: University of Bristol.

O'Neil, H. (1999). Perspectives on computer-based performance assessment of problem solving. *Computers in Human Behavior*, *15*(3), 255–268.

Pachler, N., Daly, C., Mor, Y., & Mellar, H. (2010). Formative e-assessment: Practitioner cases. *Computers & Education*, *54*(3), 715–721.

Papamitsiou, Z. K., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, *17*(4), 49–64.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Powell, S. (2014). How to increase mathematics achievement in at-risk 6th grade students (Master's Thesis). Goucher College. Retrieved from https://mdsoar.org/handle/11603/1643

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*(1), 137–150.

Raymond, C., & Usherwood, S. (2013). Assessment in simulations. *Journal of Political Science Education*, *9*(2), 157–167.

Reigeluth, C. M., & Karnopp, J. R. (2013). *Reinventing schools: It's time to break the mold*. Lanham, Maryland: R&L Education.

Learning, R. (1998). *Accelerated math*. Wisconsin Rapids, WI: author.

Rodrigues, F., & Oliveira, P. (2014). A system for formative assessment and monitoring of students' progress. *Computers & Education*, *76*, 30–41.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146.

Salvia, J., Ysseldyke, J., & Witmer, S. (2012). *Assessment: In special and inclusive education* (11th ed.). Boston: Houghton Mifflin.

Schacter, J., Herl, H., Chung, G., Dennis, R., & O'Neil, H. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior*, *15*(3), 403–418.

Schwartz, D. L., Bransford, J. D., & Sears, D. (2005). Efficiency and innovation in transfer. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 1–51). Greenwich, CT: Information Age Publishing.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M. (2016a). Advances in the science of assessment. *Educational Assessment*, *21*(1), 1–27.

Shute, V. J., & Zapata-Rivera, D. (2010). Intelligent systems. In E. Baker, P. Peterson & B. McGaw (Eds.), *Third edition of the international encyclopedia of education* (pp. 75–80). Oxford, UK: Elsevier Publishers.

Shute, V., Ke, F., & Wang, L. (in press). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Techniques to facilitate learning and motivation of serious games*. New York, NY: Springer.

Shute, V. J., Rahimi, S., & Sun, C. (in press). Measuring and supporting learning in educational games. In M. F. Young & S. T. Slota (Eds.), *Exploding the castle: Rethinking how video games & game mechanics can shape the future of education*. Inc: Information Age Publishing.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

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

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten A hog by weighing It–Or can you? evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education*, *18*(4), 289–316.

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

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

Shute, V., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and application* (pp. 535–562). Hoboken, NJ: John Wiley & Sons, Inc.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present and future. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 570–600). New York, NY: Macmillan.

Siemens, G., & Baker, R. S. J. D. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In S. B. Shum, D. Gasevic & R. Ferguson (Eds.), *Proceedings of the 2nd International conference on learning analytics and knowledge* (pp. 252–254). New York: NY, USA.

Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics—A literature review. *ICTACT Journal on Soft Computing*, *5*(4), 1,035–1,049.

Stiggins, R. (2002). Assessment for learning. *Education Week*, *21*(26), 30.

Tempelaar, D. T., Heck, A., Cuypers, H., van der Kooij, H., & van de Vrie, E. (2013). Formative assessment and learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 205–209). New York, USA: ACM.

Thelwall, M. (2000). Computer-based assessment: A versatile educational tool. *Computers & Education*, *34*(1), 37–49.

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2015). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal Early View.* doi:10.1002/berj.3215.

Tsai, F. (2013). The development and evaluation of an online formative assessment upon single-player game in E-learning environment. *Journal of Curriculum and Teaching*, *2*(2), 94–101.

Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). Advances in learning analytics and educational data mining. *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN2015)*, Bruges, Belgium. 297–306.

van der Kleij, F., Timmers, C., & Eggen, T. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. *CADMO*, *19*(1), 21–38.

van der Kleij, F. M., Eggen, T. J., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, *58*(1), 263–272.

van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511.

Wang, K. H., Wang, T., Wang, W., & Huang, S. (2006). Learning styles and formative assessment strategy: Enhancing student achievement in web-based learning. *Journal of Computer Assisted Learning*, *22*(3), 207–217.

Wang, T. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning*, *23*(3), 171–186. doi:10.1111/j.1365-2729.2006.00211.x.

Wang, T. (2010). Web-based dynamic assessment: Taking assessment as teaching and learning strategy for improving students' e-learning effectiveness. *Computers & Education*, *54*(4), 1157–1166.

Wang, T. (2011). Developing web-based assessment strategies for facilitating junior high school students to perform self-regulated learning in an e-learning environment. *Computers & Education*, *57*(2), 1801–1812.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—More than reasoning? *Intelligence*, *40*(1), 1–14.

Ysseldyke, J. E., & McLeod, S. (2007). Using technology tools to monitor response to intervention. In S. Jimerson, M. Burns & A. VanderHeyden (Eds.), *Handbook of response to intervention* (pp. 396–407). New York: Springer.

Ysseldyke, J., Spicuzza, R., Kosciolek, S., & Boys, C. (2003). Effects of a learning information system on mathematics achievement and classroom structure. *The Journal of Educational Research*, *96*(3), 163–173.

Zakrzewski, S., & Bull, J. (1998). The mass implementation and evaluation of computer-based assessments. *Assessment & Evaluation inn*, *23*(2), 141–152.