International Journal of Artificial Intelligence in Education (2021) 31:622–635
https://doi.org/10.1007/s40593-020-00198-z

COMMENTARY

Jim Greer's and Mary Mark's Reviews of Evaluation Methods for Adaptive Systems: a Brief Comment about New Goals

Benedict du Boulay[1]

Published online: 30 June 2020

Abstract

Mark and Greer's (*International Journal of Artificial Intelligence in Education, 4*(2/3), 129–153, 1993) review was very influential in setting out effective goals and methods for evaluating adaptive educational systems of all kinds. A later review brought the story up to date (Greer, *International Journal of Artificial Intelligence in Education, 26*(1), 387–392, 2016). The current paper explores a new range of evaluative goals which go beyond the quality of learning outcomes, learning efficiency, transfer, retention, and short-term motivation. While learner satisfaction has been downgraded over the years as a reliable indicator of learning quality, it cannot be wholly ignored in terms of wider issues such as the learner's developing metacognitive and meta-affective insight, regulatory competence and longer-term motivation. These factors lead on to such evaluable issues as the learner's appetite for further learning of the kind just experienced as well as for learning in general. The rise in the use of data analytics and the increasing use of AIED and computer-based learning systems in schools and universities has led to the development of orchestration systems to assist the teacher to manage their students using such systems. Orchestration systems raise new kinds of evaluation goal, such as the balance of activity, cooperation and agency between the human teacher and the adaptive systems, as well as between the learner, the systems, the teacher and, indeed, other learners. Further evaluable goals include the degree to which the teacher is alerted to the learning difficulties of the learners, the degree to which the teacher's scarce and valuable time is being used efficiently, and the degree to which the orchestration system can be used as a reflective device for teachers to examine their own practice.

Keywords Educational evaluation · Learner goals and outcomes · Teacher goals and outcomes · Metacognition · Meta-affect · Motivation · Orchestration systems

✉ Benedict du Boulay
B.du–Boulay@sussex.ac.uk

[1] Human-Centred Technology Research Group, University of Sussex, Brighton, UK

Introduction

Mark and Greer's (1993) paper was influential in that it brought together a wide range of methodologies and goals that could be applied in evaluating ITSs in particular, and AI in Education systems more generally. They described techniques for testing system designs at various stages of the design and implementation process (which they called "formative evaluations"), including fine-tuning finished systems, as well as techniques for evaluating their educational effectiveness (which they called "summative evaluations"). They referred to classic measures of educational effectiveness including the size of learning gains and/or the efficiency of learning via the amount of time to reach some criterion of success, the ability of the learner to transfer what they had learned to similar but different situations, the degree to which learners retained knowledge and skill over time and the degree to which learners were motivated to learn by their experience with the system. Their paper listed both qualitative and quantitative methods for testing and evaluation that had been used on complex software systems bearing some similarity to ITSs, such as Expert Systems. These included sensitivity analysis and pilot testing amongst others. However, they did not refer to the more general methods of experimental methodology as applied in the behavioural sciences. Clearly a contemporary paper on AIED evaluation methods would need to draw more widely on such methodologies (see, e.g. Schneider 2013, for a more recent publication on this methodology).

Their retrospective paper (Greer 2016) brought the review up to date to take account of more recently used evaluative techniques, although the evaluative goals remained largely similar:

> "using the open web and crowd sourcing to evaluate systems; comparing decisions made by adaptive learning environments and human experts; using simulated learners in the evaluation of learning environments; washing out selection bias while evaluating educational interventions; examining learning curves for evaluating systems; performing evaluations derived from mega quantities of micro measurements." (Greer 2016, page 388)

The increasing use of data analytics, mentioned by Greer in the final point of the quote above, has opened up the possibility to engage in (i) hypothesis testing in order to identify ways in which an existing system might be improved and thus the need to check whether a change in interface, pedagogy or some other adjusted feature has in fact made matters better, as well as in (ii) developing pedagogical theory. Two papers by Koedinger and his colleagues illustrate these two kinds of evaluation. In the first, Koedinger et al. (2013) analysed detailed learner logs to identify a highly specific gap in the pedagogy of their tutor which they were then able to fix. In the second, Mathan and Koedinger (2005) explored the issue of immediate vs. delayed feedback in the cognitive tutors in terms of their effects on the metacognition of learners.

The different foci of testing and developing theory mentioned above underline the multi-faceted nature of AIED as both an empirical science and an engineering

discipline. This distinction is inherited from the early days of Artificial Intelligence, where Buchanan (1988), for example, comments on these two paradigms:

> "Instead of a dichotomy of research paradigms, however, AI seems to contain a progression of steps from theorizing to engineering, from engineering to analysis, and from analysis back to theorizing. All seem important for progress." (Buchanan 1988)

The field of AIED provides a laboratory for generating and testing educational theories: for example, what is the optimal timing for different kinds of feedback? While such experiments can be conducted with human teachers, the use of systems makes the process easier, though it introduces further issues such as the degree to which one can generalise between what human teachers do effectively and what systems do effectively (see e.g., du Boulay and Luckin 2001). The fact that systems need rules of behaviour that operate at a fine level of granularity enables the generation and refinement of pedagogical principles, such as pedagogical dialogue interactions, that are not otherwise easily open to development by other means (Graesser et al. 2004).

AIED is also an engineering discipline involving the design, implementation and testing of educational systems, thus involving HCI, ergonomics, cognitive processing and interaction principles to produce well-constructed systems that help learners achieve educational goals or help teachers manage those learners or indeed both (see e.g., Kirschner et al. 2011; Mayer 2014).

Educational Effectiveness

In our work at Sussex we have made extensive use of the evaluative technique from empirical science, mentioned by Mark & Greer, of using learning gains to compare two versions of the same system that differ in a *single* aspect. Typically this has been done in order to test a particular educational principle: for example, how effective is matching vs. mismatching the goal orientation of a learner to the degree to which the system reacts appropriately to that goal orientation (du Boulay 2011). For a more general review of the different kinds of adaptation available in AIED systems and their evaluations, see Aleven et al. (2017).

Unlike comparisons between two systems that differ in a single aspect, there have been many comparisons of educational interventions that differ in many characteristics. For example, over the last decade there have been at least nine meta-reviews and meta-analyses that have compared the educational effectiveness of adaptive systems for a topic versus human teachers of that topic, working either with whole classes or one-to-one (see, e.g. Ma et al. 2014; VanLehn 2011). For an overview of these nine meta-reviews and meta-analyses, see du Boulay (2016).

One example of comparing an adaptive system against human teachers involved the Cognitive Tutor for Algebra. A large, complex and carefully organised, multi-state experiment was conducted in the USA over two years (Pane et al. 2014). The main outcome measures compared pre/post results of learners in matched schools, half continuing to teach as they had before and the other half incorporating the intelligent tutor in a blended manner. In the first year there were no significant differences, but by

the second year, significant differences did occur, no doubt as the teachers had figured out how best to orchestrate the use of the tutoring system. A later paper detailed an analysis of the effects on the teachers of taking on and adapting to the new blended role (Karam et al. 2017). They found that:

> ". . . teachers implemented the blended curriculum with low fidelity. Teachers had most difficulty allocating the recommended amount of time for the math lab and content. The study also found that the blended-curriculum teachers in the second year reverted to more traditional approach to instruction and spent less time on inquiry based instruction than in the first year, although they continued to use this approach at a higher level than teachers in the control schools. The study findings suggest that teacher adjustment of instruction in the second year, specifically balancing the amount of traditional instruction with inquiry instruction, in combination to the use of the math software contributed to the performance of the program." (Karam et al. 2017, page 399).

The current paper is concerned with the issue of the educational effectiveness of AIED systems and augments the work of Mark and Greer by identifying a number of evaluable educational goals and outcomes not mentioned in their work, such as paying attention to the effects on teachers of an intervention involving new technology.

Following the increasing use of AIED and other computer-based learning systems in schools, "in the wild", as well as the increasing sophistication of learning analytic methods, has led to both the need for, and the capability to build, systems to help teachers manage classes in such schools. Thus, one trend that has become more prominent over the last few years is the development of blended and orchestration systems (see, e.g. Dillenbourg 2013). These systems assume that there will be a human teacher in the loop and that the learner will be exposed to individual or small group work with an AIED system, individual or small group interaction with a human teacher as well as whole class teaching from the human teacher, possibly assisted by public use of the AIED system. Such systems open up a range of new educational goals for evaluation, e.g. the degree to which the human teacher's time is being used effectively.

A second trend has been to take better account of the non-cognitive aspects of learners including their desire (or not) to learn, their longer-term motivation and values, their affective trajectory before, during and after learning, and their metacognition and meta-affect. Table 1 lists a number of broad areas of evaluable goals for learners and Table 2 does the same for teachers. The rows a-g in Table 1's evaluable goals are those already covered by Mark & Greer.

A third trend, that supports the first two, is the increasing use of AIED systems in schools and universities as well as the rise of educational data-mining and learner analytics. These provide tools and methods to support both empirical science questions concerning the nature of learning and teaching as well as design engineering issues around improving learning interactions with systems.

Table 1 Evaluable goals focusing on learners

1. Learner-Focused
   a. Skill or concept outcome measure, e.g. via standard pre-/post- test designs.
   b. Skill or concept learning efficiency, e.g. via time to achieve criterion, learning curves.
   c. Satisfaction with learning experience, e.g. via qualitative surveys or interviews.
   d. Satisfaction with learning outcome, e.g. via qualitative surveys or interviews.
   e. Willingness to keep participating to the end of the course, e.g. via dropout rates (where applicable), in other words short term motivation.
   f. Retention of new skills and knowledge, e.g. via delayed post-tests.
   g. Ability to transfer skills and understanding to new but related areas, e.g. via posttests.
   h. Meta-cognitive understanding of the learning outcome, e.g. via think aloud problem-solving, observation of future learning.
   i. Metacognitive regulatory skill, e.g. via think aloud problem-solving, observation of future learning.
   j. Increased meta-affective awareness feelings while learning, e.g. via self-reports.
   k. Increased capability to regulate feelings associated with learning, e.g. via refection via open learner models.
   l. Increased pleasure in the act of learning for its own sake, e.g. via long-delayed post-questionnaires or interviews.
   m. Increased participation in learning activities of all kinds, e.g. via long-term cohort studies.
   n. Increased and/or more effective cooperation between learners.

Learner-Focused Goals

There has been increasing work on identifying the affective trajectories of learners and designing systems to manage these trajectories in order to improve learning gains. Systems adapt their scaffolding, feedback, help and task selection (see, e.g. Arroyo et al. 2014). Such adaptations attempt (i) to maximise the chances that learners will enter, and remain in, productive affective states and (ii) minimise the chances that they will enter, or fail to exit from, non-productive affective learning states.

Motivation is a complex notion that includes aspects of cognition, meta-cognition, affect, meta-affect and values (Schunk et al. 2008). A learner's relatively transient feelings and self-efficacy, their interpretation of those internal states and the consequent cognitive and motivational changes affect not just their expectations for the future but also their interpretation of their current situation and even their understanding of past learning experiences. Examining the nature of self-efficacy, McQuiggan et al. (2008) built a *dynamic* predictive model of self-efficacy based on pre-test data as well as physiological data gathered during learning. In addition, Bernacki et al. (2015) explored how learners' self-efficacy judgments varied even over the period of a single problem-solving session and found that:

> "Their prior performance (i.e., accuracy) predicted subsequent self-efficacy judgments, but this relationship diminished over time as judgments were decreasingly informed by accuracy and increasingly informed by fluency." (Bernacki et al. 2015, page 99))

In earlier papers, I have tried to characterise some of this motivational complexity within an AIED context (du Boulay et al. 2010) and reviewed AIED work in this area (du Boulay 2018). While AIED has been concerned with motivational issues for many

years (see, e.g. del Soldato and du Boulay 1995), there has been much recent work on many aspects of it.

Quite a lot of this work has been concerned with identifying and managing the interwoven cognitive, emotional, and motivational trajectories *during* instruction and evaluating immediate learning outcomes (see, e.g. Arroyo et al. 2014). For example, Arroyo and her colleagues' system included feedback to learners based on the "growth mindset" (Dweck 1999), but (as far as I know) there has not been an analysis of the degree to which this mindset was retained or transferred to other educational contexts.

Clearly, an important factor in motivating learners is capturing their interest, as a stimulus before and within a lesson as well as, potentially, even after it has finished. For example, Harackiewicz et al. (2016) argue that:

> "Interest is a powerful motivational process that energizes learning, guides academic and career trajectories, and is essential to academic success. Interest is both a psychological state of attention and affect toward a particular object or topic, and *an enduring predisposition* to *reengage over time* [my emphasis]." (Harackiewicz et al. 2016, page 220)

There are various ways to capture interest, including exploiting the social aspects of learning as well as trying to make the material to be learned more obviously relevant to the learners.

An example of capturing a social aspect of the learner's interest is provided by Kelly et al. (2013a). They found that including videos of the students' own human teacher providing motivational feedback within an AIED system was more effective than using an animated pedagogical agent, and that this improved homework completion rates too. Taking account of the importance of the social aspects of learning has also been reported by Olsen et al. (2019). They compared students learning fractions with an AIED system either individually, collaboratively or with a mixture of both modes. The students had the best learning outcomes in the combined condition as compared to working wholly individually or wholly collaboratively.

An example of capturing interest through relevancy is provided by Walkington and Bernack (2019) who found that adjusting the context of algebra problems to take account of the student's out of school interests was beneficial. Finally, Klebanov et al. (2017) found that engaging students in experimental writing helped them reflect on and begin to understand the "utility value" for them of the STEM subjects they were studying.

There is potentially a tension between capturing interest and provoking engagement as against fostering effective learning, in that interest and engagement are (mostly) necessary but not sufficient conditions for learning. One reason for this is that learning (other than learning by rote) also needs a reflective component, and this can get drowned out if engagement and fun do not leave enough room for it. This tension is much in evidence in the use of games in education, where the main argument for their use is based on their ability to engage. However, individual differences in self-regulation ability and cognitive load capacity can affect how much is actually learned in a game-based learning environment (for reviews of the positive and some of the negative attributes of game-based learning see, e.g. Vlachopoulos and Makri 2017;

Zhonggen 2019). Even though there are many reports of the effectiveness of games in education, many questions remain unresolved (de Freitas 2018).

There have been many systems that have engaged learners at the metacognitive level (see, e.g. Azevedo and Aleven 2013). Typically, the metacognitive aspects of the interaction have been aimed to improve learning outcomes, but there have been some systems where the specific aim was to increase metacognitive awareness and regulation as an end in itself (see, e.g. Azevedo et al. 2009). In a similar vein, Long et al. (2015) designed a tutor to help students *learn how to learn* in a problem-solving session by teaching them how to select an appropriate next problem that conforms to the "Mastery Rule", namely that the next problem should require skills that have already been learnt as well as at least one that needs more practice.

Evaluating such systems requires an examination of whether the specific metacognitive skills being taught survive to be deployed by the learner in future learning with a similar system (retention), or even better, whether they survive to be deployed in dissimilar learning situations (transfer).

Many teachers aim not only to teach some specific set of skills or some specific understanding, but also hope that their learners will further develop their more general desire to learn (see, e.g. Maehr 2012). The teacher hopes that the experience of getting to understand something, or the ability to exercise a new skill, will be pleasurable and memorable in itself, and so will act as an intrinsic motivational force towards engaging in further learning experiences. Ideally one would like the learner to be aware of such learning and its pleasure, but even a relatively unreflective pleasure would be beneficial. Moving to a more reflective awareness may require assistance from the teacher to help learners develop their "meta-affection" and "meta-motivation", i.e. their understanding and regulation of their own affective and motivational processes.

Evaluating the above longer-term outcome means that, in the short term, it is necessary to augment any testing of the satisfaction of the learner with their experience of the learning process by also checking (and possibly contrasting) their degree of understanding of, and satisfaction with, the learning outcome, namely their increase in skill or understanding. It might be that they had a pleasant time, but did not learn much, or a poor time and learned much, as well as the other two possibilities. Such meta-affective and meta-cognitive mentoring needs to be sensitive to the potential for learners to equate a pleasant learning experience with an effective learning outcome, or a challenging learning experience with an ineffective outcome (see, e.g. Whitelock and Scanlon 1996). It also needs to be sensitive to the possibility that learners may prefer a more passive learning experience to a more active one, despite the latter's generally greater educational effectiveness (Deslauriers et al. 2019). In the longer term, one would also need to track their future learning choices and demeanours, somewhat in the same way as using a delayed post-test to see what proportion of any initial learning gain had persisted.

The rise in the availability of MOOCs has raised another educational issue, that of dropout rates (Liyanagunawardena et al. 2013). A new goal for such systems is measured by the proportion of learners who work their way through all of the available lessons and tasks, to some extent irrespective the quality of the learning or indeed its efficiency. Various attempts have been made to try to improve retention. These include,

Table 2 Evaluable goals focusing on teachers

2. Teacher-Focused
a. Change in the division of labour and use of time between the teacher and whatever other adaptive systems are being used, e.g. via class observation and interviews with the teacher and with learners.
b. Ability to diagnose better individual learner and/or class difficulties, e.g. via post class interviews with teacher.
c. Ability to fix better individual learner and/or class difficulties, e.g. via observation of teacher in class.
d. More productive use of (scarce) teacher time, e.g. via class observation of teacher in class.
e. Reduction in non-productive teacher tasks, e.g. via class observation of teacher in class.
f. Satisfaction with the teaching experience, e.g. via qualitative surveys or interviews.
g. Satisfaction with the teaching outcome, e.g. via qualitative surveys or interviews.
h. Reification of orchestration itself allowing reflection by the teacher, e.g. via post-hoc interviews

for example (i) embedding AIED components within the MOOC (Aleven et al. 2016), (ii) trying to identify, and then build on, features that are most predictive of retention, such as learner engagement (Bakki et al. 2015; Deng et al. 2020; Joksimovic et al. 2018) and (iii) exploring students' perceptions of the effectiveness of the course and the quality of interaction with the tutor (Hone and El Said 2016). For a recent overview of research on MOOCs, see Deng et al. (2019). They identified five important issues:

> "(1) evidence-based research on non-mainstream consumers of MOOCs is scarce; (2) the role of learner factors is oversimplified in evidence-based MOOC research; (3) there is no attempt to reconcile different approaches to measuring learner engagement with MOOCs; (4) measures of learning outcomes lack sophistication and are often based on single variables; and (5) the relationships between many of the key learning and teaching factors have not been clarified." (Deng et al. 2019, page 48)

Teacher-Focused Goals

AIED research has largely focused on assisting learners rather than on assisting teachers, although there has been a thread of ongoing work in the latter area, both in terms of an analysis of the role of the teacher when AIED systems are deployed (see, e.g. Vivet 1992) as well as in terms of systems designed to help the teacher. For example, in the latter case, Yacef (2002) set out a number of roles for "intelligent teaching assistant systems" as follows and described her own and others' work in this area:

> "Help in diagnosis and assessment of learning . . .
> Help in generating tailored material for a particular student . . .
> Help for monitoring one student during the execution of an exercise . . .
> Help for analysing or synthesising results . . .
> Help in creating/defining the ITS . . .

> Reducing the quantity or length of burdensome tasks that can be automated or facilitated . . .
> Improving the quality of the teaching process, by providing new or better tools and feedback to the teacher . . .” (Yacef 2002, pages 136–7)

The Changing Role of the Teacher

The Introduction to this paper has already mentioned the analysis of the teachers’ roles in the large-scale evaluation of the Cogntive Tutor for Algebra where the “teachers implemented the blended curriculum with low fidelity” and the effect that this had on learning outcomes (Karam et al. 2017).

An early paper in this area examined why students seemed to prefer help from an AIED system compared to help from teacher (Schofield et al. 1994). One reason for this seemed to be that introducing the AIED system as a kind of classroom assistant, freed the teacher to provide more individualised assistance. The combination of the extra resource provided by the system together with the more targeted assistance from the teacher led both to better learning outcomes as well as to more satisfaction for the learners. In a later paper, the same author pointed out the subtle effects for the teacher that introducing computers into a class produces, not least on the mode of teaching that teachers adopt (Schofield 1997).

More recently, there has been a detailed analysis of the different ways that the triad of learner, teacher and AIED system may interact. Kessler et al. (2019) observed teachers taking different roles when an AIED system for mathematics was deployed. These included, among others, the teacher delegating the teaching to the system, the teacher facilitating the learner’s use of the system and the teacher facilitating the learner’s understanding of the mathematics in the system, as well as the teacher directly interacting with the learner independently of the learner’s interaction with the system. The study plotted both the interaction roles and the learner outcomes.

Indeed, failure to acknowledge the importance of the human teacher in the loop has caused various problems, such as high rates of student dissatisfaction (Tabor 2018).

Orchestration Systems

The realisation of the *centrality* of the human teacher in the educational ecosystem that now also includes AIED systems has led to the development of various kinds of “orchestration” system (Dillenbourg 2013) to assist the teacher manage the added complexity of having AIED systems in their classes, as anticipated by Yacef (2002), above.

These orchestration systems fall into three broad types. There are systems that are designed to be used in a situation where all the learners are working with an AIED system, and the teacher needs help in making best use of her time to provide extra help to just those learners who need it most (Holstein et al. 2018, 2019). There are systems that help the teacher track and monitor learners, or groups of learners, using more standard rather than AIED learning technology (Cheema et al. 2016). Finally there are systems that play a dual role of working directly with learners but also offering the

teacher a dashboard that indicates general difficulties that learners are having which can then be addressed by the teacher in a whole-class mode (Heffernan and Heffernan 2014). For example, Kelly et al. (2013a, b) describes the use of the ASSISTments system that students use to do their homework which also provides, the next day, an analysis for the teacher of their common and individual difficulties to help the teacher choose what issues to focus on.

In a series of papers, Martinez-Maldonado and his colleagues have developed a system that can provide guidance and feedback to the teacher in a classroom where the students are working in small groups at interactive tabletops (Martinez-Maldonado et al. 2013, 2015, 2016, 2018). In this case, the classroom situation is more complex than that for Holstein et al.'s orchestration system described above, as the students are more mobile within the class, working collaboratively and not all their learning activities can be directly logged by the tabletops.

In addition to the learner-focused goals already mentioned, orchestration systems bring with them a new set of evaluable goals relating to the teacher. Such systems were not yet built when Mark and Greer (1993) wrote their paper, and were not yet strongly in evidence when Greer (2016) wrote his retrospective commentary on their earlier paper.

One such goal revolves around the potential change in the division of labour between the adaptive system or e-learning system and the human teacher. For example, is the teacher's role and use of time with the learners changed as a result? A related goal is do they enable the teacher to use her time in class more effectively becuase the system helps to identify those individuals who need the extra human help the most (Holstein et al. 2018)? Another goal is to what degree can teachers save their own and class time by having the learners using teaching component of the system for homework, and then have the system identify issues of concern to majority of the learners for the teacher to deal with in class (Roschelle et al. 2016)? A further goal is do such systems enable the teacher to reflect on any differences between their planned and actual orchestration with the possibility of more effective orchestration in the future?

These goals can be evaluated by post hoc interviews with teachers as well as by a comparative analysis of the use of teachers' time working with and without the advice about which learner or what topic to concentrate on. Of course, it does not necessarily follow that a more focused use of the teachers' time with those who need help will always lead to better learning outcomes for all the learners, but it is a very reasonable hypothesis and has been shown to be the case (see, e.g. Roschelle et al. 2016).

Conclusions

This paper has acknowledged and built on the earlier work of Greer and Mark in identifying evaluation methods and evaluation goals as applied to AIED systems. Their distinction between formative and summative evaluation methods has been recast in terms of the dual nature of AIED as both a design engineering discipline for building interactive educational systems and an empirical science of developing theory in learning and teaching.

The paper has plotted trends in the development of AIED to identify, not new methods of evaluation, but new goals for educational evaluation. These have broadly

divided into two kinds. There are goals now focusing on the learner as a feeling and thinking being, their learning experience more broadly, retention (in the case of MOOCs), their insight into their own learning, and their motivation to undertake future learning.

In parallel, there are also goals focusing on the role of the teacher in relation to the deployment of AIED systems, the teachers' experience of the classroom, their efficiency and satisfaction. There are also systems to assist the teacher in various ways, e.g. identifying which topics their students need personal help with, or which parts of the homework were largely done well and so need little further feedback.

Underpinning these changes is the increase in use of AIED systems in schools and universities and the role of data-mining and learner analytics that enable the design and execution of analyses to assist AIED as a design discipline as well as an empirical science.

Acknowledgements I thank Charlotte Robinson, Emeline Brule and Grazia Ragone and other members of the Human-Centred Technology Group at Sussex for helpful comments on an earlier draft of this paper. I also thank the referees of the paper for their insightful and motivating remarks as well as for pointing me to papers that I have found stimulating and useful.

References

Aleven, V., Sewall, J., Popescu, O., Ringenberg, M., van Velsen, M., & Demi, S. (2016). Embedding intelligent tutoring systems in MOOCs and e-learning platforms. In A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Proceedings of the 13th international conference on intelligent tutoring systems, ITS 2016. Lecture notes in computer science, vol 9684* (pp. 409–415). Springer.

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 522–560). New York: Routledge.

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

Azevedo, R., & Aleven, V. (2013). *International handbook of metacognition and learning technologies*. Springer.

Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C., & Fike, A. (2009). *MetaTutor: a metacognitive tool for enhancing self-regulated learning*. Paper presented at the AAAI fall symposium (FS-09-02).

Bakki, A., Oubahssi, L., Cherkaoui, C., & George, S. (2015). Motivation and engagement in MOOCs: How to increase learning motivation by adapting pedagogical scenarios? In *EC-TEL 2015 tenth European conference on technology enhanced learning: Design for Teaching in a networked world* (pp. 556–559). Toledo: Spinger.

Bernacki, M. L., Nokes-Malach, T., & Aleven, V. (2015). Examining self-efficacy during learning: Variability and relations to behavior, performance, and learning. *Metacognition and Learning, 10*(1), 99–117. www.doi.org/10.1007/s11409-014-9127-x.

Buchanan, B. G. (1988). Artificial Intelligence as an Experimental Science. In J. H. Fetzer (Ed.), *Aspects of artificial intelligence* (pp. 209–250). Dordrecht: Kluwer Academic Publishers.

Cheema, S., VanLehn, K., Burkhardt, H., Pead, D., & Schoenfeld, A. (2016). *Electronic posters to support formative assessment*. Paper presented at the CHI EA '16 Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems San Jose, CA, USA.

de Freitas, S. (2018). Are games effective learning tools? A review of educational games. *Educational Technology & Society, 21*(2), 74–84.

del Soldato, T., & du Boulay, B. (1995). Implementation of motivational tactics in tutoring systems. *International Journal of Artificial Intelligence in Education, 6*(4), 337–378.

Deng, R., Benckendorff, P., & Gannaway, D. (2019). Progress and new directions for teaching and learning in MOOCs. *Computers & Education, 129*, 48–60. www.doi.org/10.1016/j.compedu.2018.10.019.

Deng, R., Benckendorff, P., & Gannaway, D. (2020). Learner engagement in MOOCs: Scale development and validation. *British Journal of Educational Technology, 51*(1), 245–262. https://doi.org/10.1111/bjet.12810.

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences of the United States of America, 116*(39), 19251–19257. https://doi.org/10.1073/pnas.1821936116.

Dillenbourg, P. (2013). Design for classroom orchestration. *Computers & Education, 69*, 485–492. https://doi.org/10.1016/j.compedu.2013.04.013.

du Boulay, B. (2011). Motivationally intelligent educational systems: The contribution of the human Centred technology research group. *Technology, Instruction, Cognition and Learning, 8*(3–4), 229–254.

du Boulay, B. (2016). Artificial intelligence as an effective classroom assistant. *IEEE Intelligent Systems, 31*(6), 76–81.

du Boulay, B. (2018). Intelligent tutoring systems that adapt to learner motivation. In S. D. Craig (Ed.), *Tutoring and intelligent tutoring systems* (pp. 103–128). New York: Nova Science Publishers, Inc..

du Boulay, B., & Luckin, R. (2001). The plausibility problem: An initial analysis. In M. Beynon, C. L. Nehaniv, & K. Dautenhahn (Eds.), *4th international conference on cognitive technology - instruments of mind, Coventry, England* (pp. 289–300).

du Boulay, B., Avramides, K., Luckin, R., Martinez-Miron, E., Rebolledo-Mendez, G., & Carr, A. (2010). Towards systems that care: A conceptual framework based on motivation, metacognition and affect. *International Journal of Artificial Intelligence in Education, 20*(3), 197–229.

Dweck, C. S. (1999). *Self-theories: Their role in personality, motivation, and development*. Philadelphia: Psychology Press.

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

Greer, J. (2016). Evaluation methods for intelligent tutoring systems revisited. *International Journal of Artificial Intelligence in Education, 26*(1), 387–392. https://doi.org/10.1007/s40593-015-0043-2.

Harackiewicz, J. M., Smith, J. L., & Priniski, S. J. (2016). Interest matters: The importance of promoting interest in education. *Policy Insights from the Behavioral and Brain Sciences, 3*(2), 220–227. www.doi.org/10.1177/2372732216655542.

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. https://doi.org/10.1007/s40593-014-0024-x.

Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In C. P. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Artificial intelligence in education: 19th international conference, AIED 2018, London, UK, June 27–30, 2018 proceedings, part I* (pp. 154–168). Cham: Springer.

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learner Analytics, 6*(2), 27–52. https://doi.org/10.18608/jla.2019.62.3.

Hone, K. S., & El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. *Computers & Education, 98*, 157–168. https://doi.org/10.1016/j.compedu.2016.03.016.

Joksimovic, S., Poquet, O., Kovanovic, V., Kovanovic, V., Dowell, N., Mills, C., et al. (2018). How do we model learning at scale? A systematic review of research on MOOCs. *Review of Educational Research, 88*(1), 43–86. https://doi.org/10.3102/0034654317740335.

Karam, R., Pane, J. F., Griffin, B. A., Robyn, A., Phillips, A., & Daugherty, L. (2017). Examining the implementation of technology-based blended algebra I curriculum at scale. *Educational Technology Research & Development, 65*, 399–425. https://doi.org/10.1007/s11423-016-9498-6.

Kelly, K. M., Heffernan, N. T., D'Mello, S., Namais, J., & Strain, A. C. (2013a). Adding teacher-created motivational video to an ITS. Paper presented at the Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference.

Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, J., & Goldstein, D. S. (2013b). Estimating the effect of web-based homework. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *16th International Conference, AIED 2013, , July 9–13, 2013. Proceedings* (pp. 824–827). Springer.

Kessler, A., Boston, M., & Stein, M. K. (2019). Exploring how teachers support students' mathematical learning in computer-directed learning environments. *Information and Learning Sciences, 121*(1/2), 52–78. https://doi.org/10.1108/ILS-07-2019-0075.

Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior, 27*(1), 99–105.

Klebanov, B. B., Burstein, J., Harackiewicz, J. M., Priniski, S. J., & Mulholland, M. (2017). Reflective writing about the utility value of science as a tool for increasing STEM motivation and retention – can AI help scale up? *International Journal of Artificial Intelligence in Education, 27*, 791–818. www.doi.org/10.1007/s40593-017-0141-4.

Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better student models to improve student learning. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education: 16th international conference, AIED 2013* (pp. 421–430). Memphis: Springer.

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008-2012. *International Review of Research in Open and Distance Learning, 14*(3), 202–227.

Long, Y., Aman, Z., & Aleven, V. (2015). Motivational design in an intelligent tutoring system that helps students make good task selection decisions. In C. Conati, N. Heffernan, A. Mitrovic, & M. Verdejo (Eds.), *Artificial Intelligence in Education. AIED 2015* (pp. 226–236). Springer.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology, 106*(4), 901–918. https://doi.org/10.1037/a0037123.

Maehr, M. L. (2012). *Encouraging a continuing personal Investment in Learning: Motivation as an instructional outcome*. Charlotte: Information Age Publishing.

Mark, M. A., & Greer, J. E. (1993). Evaluation methodologies for intelligent tutoring systems. *International Journal of Artificial Intelligence in Education, 4*(2/3), 129–153.

Martinez-Maldonado, R., Dimitriadis, Y., Kay, J., Yacef, K., & Edbauer, M.-T. (2013). MTClassroom and MTDashboard: Supporting analysis of teacher attention in an orchestrated multi-tabletop classroom. Paper presented at the international conference on computer supported collaborative learning (CSCL2013), Madison, Wisconsin, USA.

Martinez-Maldonado, R., Clayphan, A., Yacef, K., & Kay, J. (2015). MTFeedback: Providing notifications to enhance teacher awareness of small group work in the classroom. *IEEE Transactions on Learning Technologies, 8*(2), 187–200. www.doi.org/10.1109/TLT.2014.2365027.

Martinez-Maldonado, R., Schneider, B., Charleer, S., Shum, S. B., Klerkx, J., & Duval, E. (2016). *Interactive surfaces and learning analytics: Data, orchestration aspects, pedagogical uses and challenges*. Paper presented at the International Learning Analytics & Knowledge Conference, Edinburgh, UK.

Martinez-Maldonado, R., Echeverria, V., Santos, O. C., Santos, A. D. P. D., & Yacef, K. (2018). *Physical Learning Analytics: A Multimodal Perspective*. Paper presented at the LAK'18 proceedings of the 8th international conference on learning analytics and knowledge, Sydney.

Mathan, S., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational psychologist, 40*(4), 257–265. www.doi.org/10.1207/s15326985ep4004_7.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge University Press.

McQuiggan, S. w., Mott, B. W., & Lester, J. C. (2008). Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction, 18*, 81–123. https://doi.org/10.1007/s11257-007-9040-y.

Olsen, J. K., Rummel, N., & Aleven, V. (2019). It is not either or: An initial investigation into combining collaborative and individual learning using an ITS. *International Journal of Computer-Supported Collaborative Learning, 14*, 353–381. https://doi.org/10.1007/s11412-019-09307-0.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis, 36*(2), 127–144. https://doi.org/10.3102/0162373713507480.

Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open, 2*(4), 1–12. https://doi.org/10.1177/2332858416673968.

Schneider, S. (2013). *Experimental design in the behavioral and social sciences*. London: Sage Publishing.

Schofield, J. W. (1997). Computers and classroom social processes — A review of the literature. *Social Science Computer Review, 15*(1), 27–39 https://doi-org.ezproxy.sussex.ac.uk/10.1177/089443939701500104.

Schofield, J. W., Eurich-Fulcer, R., & Britt, C. L. (1994). Teachers, computer tutors, and teaching: The artificially intelligent tutor as an agent for classroom change. *American Educational Research Journal, 31*(3), 579–607. www.doi.org/10.2307/1163227.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research and applications* (3rd ed.). Upper Saddle River: Pearson, Merrill, Prentice Hall.

Tabor, N. (2018). Brooklyn students are protesting Silicon Valley's favorite education program. *Intelligencer*.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369.

Vivet, M. (1992). Uses of ITS: Which role for the teacher? In E. Costa (Ed.), *New directions for intelligent tutoring systems* (pp. 171–180). Berlin: Springer.

Vlachopoulos, D., & Makri, A. (2017). The effect of games and simulations on higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education, 14*(22). https://doi.org/10.1186/s41239-017-0062-1.

Walkington, C., & Bernacki, M. L. (2019). Personalizing algebra to Students' individual interests in an intelligent tutoring system: Moderators of impact. *International Journal of Artificial Intelligence in Education, 29*(2), 58–88. https://doi.org/10.1007/s40593-018-0168-1.

Whitelock, D., & Scanlon, E. (1996). Motivation, media and motion: Reviewing a computer supported collaborative learning experience. In P. Brna, A. Paiva, & J. Self (Eds.), *European conference on artificial intelligence in education* (pp. 276–283). Lisbon: Fundacao Calouste Gulbenkian.

Yacef, K. (2002). *Intelligent teaching assistant systems*. Paper presented at the ICCE '02: Proceedings of the international conference on computers in education.

Zhonggen, Y. (2019). A meta-analysis of use of serious games in education over a decade. *International Journal of Computer Games Technology., 2019*, 1–8. https://doi.org/10.1155/2019/4797032.