Building a Formative Assessment System That is Easy to Adopt Yet Supports Long-term Improvement: A Review of the Literature and Design Recommendations

Ellis Solaiman, Joshua Barney and Martin Smith
School of Computing Science, Newcastle University, Newcastle upon Tyne, U.K.

Keywords: Formative Assessment, Assessment Design, Review, Feedback Automation, Education Technology.

Abstract: Formative assessment has been shown to be an effective teaching tool, yet is infrequently used in practice. With the intent of building a formative e-assessment platform, we examine research on formative practices and supporting computer-based systems with a focus on: institutional barriers to adoption of previous systems; senses in which students and teachers can improve their practices across varying timescales; and collectible data (self-reported or otherwise) necessary or advantageous in supporting these processes. From this research we identify the minimal set of data which adequately supports these processes of improvement, arrive at a set of requirements and recommendations for an innovative system which collects, processes, and presents this data appropriately, and from these requirements design the architecture of an extensible electronic formative assessment system which balances the need for complex long-term analytics with that of accessibility.

1 INTRODUCTION

Formative assessment refers to the use of testing to allow both students to reflect on their own learning and teachers to determine the needs of individual students, or groups of students. This can also be described as *assessment for learning*. Formative assessment is often contrasted with summative assessment, described as *assessment of learning*, which uses testing as a means for summarizing the performance of students and their level of knowledge at the end of a course (Looney, 2011) (Black and Wiliam, 2009). Despite the clear benefits of formative assessment as an effective teaching tool, it is infrequently used in practice. Also whilst VLEs (Virtual Learning Environments) are widespread in higher education, their formative abilities are often severely underutilized (Blin and Munro, 2008). Even when change is not resisted by faculty staff, adoption of educational technology tends to proceed at a slow pace, and often innovation is driven by lone individuals deciding to use technologies without backing from above (Russell, 2009). Although perception of student expectation can provide positive incentive, technical barriers along with a lack of institutional support are key suppressing factors (King and Boyatt, 2015). Currently available Education technologies are hampered by technical problems such as lack of integration with systems in use within institutions, and educational data-mining efforts fail due to the absence of tools which are easy for non-experts to use (Romero and Ventura, 2010). A key factor inhibiting the adoption of formative assessment technologies is staff time. Teacher time is an especially scarce resource – those who see the value of educational technology may still avoid full utilization for this reason; *"These are all great initiatives, but I'm running out of hours in a day ... I'm in overload!"* (Hall and Hall, 2010). Research by (Macfadyen and Dawson, 2012) highlights that in order to overcome individual and group resistance to innovation and change, planning processes must create conditions that allow participants to both think and feel positively about change; *"conditions that appeal to both the heart and the head"*. Without investing resources for effective automated data analysis and visualization, large quantities of data will not compel social or cultural change (Kotter and Cohen, 2002); *"increasing observability of change is an effective strategy for reducing resistance"* (Rogers, 1995).

This paper builds upon and complements previous work and literature reviews such as (Gikandi et al., 2011), by taking a more integrative approach. We attempt through the development a holistic theoretic framework,to arrive at the design of a system suitable for active use both as a tool for student and teacher

Solaiman, E., Barney, J. and Smith, M.
Building a Formative Assessment System That is Easy to Adopt Yet Supports Long-term Improvement: A Review of the Literature and Design Recommendations.
DOI: 10.5220/0006339404420449
In *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017) - Volume 1*, pages 442-449
ISBN: 978-989-758-239-4

improvement, and as a platform for further empirical research about the impact of various formative strategies. Initially, multiple high impact journals in the field of educational technology were identified, and the titles and abstract of all publications from the the last 5 years were manually examined. A number of potentially relevant papers were identified and scrutinized in further detail, and common subjects and references were noted. Keywords related to these subjects were used to find further papers using academic search engines. The process of reading papers, discovered through references or search, continued recursively until a clear picture of a range of formative and learning practices emerged. A conceptual framework was developed to organize these practices by time and individual, and the body of collected papers was reexamined in the context of this framework to explore the data-collection requirements of each of the identified practices. Recommendations are made for implementation, and from these recommendations, the design of an extensible system to collect this data is presented.

2 CLASSIFICATION OF FEEDBACK BY USER AND DURATION

Multiple researchers distinguish types of formative assessment based on their duration. Allal and Schwartz (Allal and Schwartz, 1996) use the terminology of "Level 1" and "Level 2' formative assessments, which directly benefit students who are assessed as or use data gathered to benefit future instructional activities or new groups of students respectively. Alternatively, (Wiliam, 2006) distinguishes between short-, medium-, and long-cycle formative assessment, which operate within and between lessons, within and between teaching units, and across marking periods, semesters or even years. In addition to acting as a helpful framework within which to classify and organize research, this viewpoint is useful as a design tool – collecting data required for longer-term formative processes requires action over a sustained period of time, and this affects the short-term experience of new users. As discussed below, a system must show immediate value in order to have the best chance at being widely adopted, and balancing the conflict between supporting longer-term goals and maintaining accessibility is a design challenge that deserves thinking about explicitly. In addition to distinguishing processes by time we also do so by the agents involved (See Figure 1). Our framework differs from previous formulations like that of (Wiliam and Thompson, 2007) in that we conceive of teachers not just in terms of the role they can play in providing feedback to students but also as individuals who can benefit from feedback themselves.

3 FEEDBACK PROCESSES AND THEIR DATA COLLECTION REQUIREMENTS

We examine the specifics of feedback processes below, considering issues related by their similar data-requirements together. This view allows us to move closer to the design of a real system.

3.1 Formative Feedback Principles

Synthesizing research literature, (Nicol and Macfarlane-Dick, 2006) present seven principles, reordered for clarity here, arguing effective formative feedback practice:

1. delivers high quality information to students about their learning. 2. helps clarify what good performance is (goals, criteria, expected standards). 3. provides opportunities to close the gap between current and desired performance. 4. facilitates the development of self-assessment (reflection) in learning. 5. encourages positive motivational beliefs and self-esteem. 6. encourages teacher and peer dialogue around learning. 7. provides information to teachers that can be used to help shape teaching.

The first three principles mirror (Wiliam and Thompson, 2007)'s theory of formative assessment, which emphasizes establishing *where the learners are in their learning*, *where they are going*, and *what needs to be done to get them there*. Black and William, (Black and Wiliam, 1998) warn that classroom cultures focused on rewards or grades encourage students to game the system, avoid difficult tasks, and where difficulties or poor results are encountered, *"retire hurt"*. By 2004, this had been strengthened in (Black et al., 2004) into an explicit recommendation to adopt comment-only marking practices; *"Students given marks are likely to see it as a way to compare themselves with others; those given only comments see it as helping them to improve. The latter group outperforms the former."* A literature review by (Gikandi et al., 2011) provides supporting studies and explicitly reaffirms the above seven principles as *"an essential condition for effective formative feedback"*.

Whilst the above principles may be sufficient to guide the behavior of teachers in conversation with

	Short-term	Medium-term	Long-term
Students	Checking understanding of subject, identifying own misconceptions.	Understanding progress and thinking about own learning (meta-cognition).	Subject mastery, perhaps through spaced repetition.
Teachers	Identifying common misconceptions in the student body, planning next lesson.	Understanding broad proficiencies of class across the whole syllabus.	Constructing next year's course materials, and designing new summative assessments.

Figure 1: Conceptual framework of feedback for students and teachers in the short, medium, and long terms.

students in a physical classroom, there are pragmatic details to consider for their implementation within computer-based systems. For example Hattie and Timperley (Hattie and Timperley, 2007) distinguish between task-orientated feedback such as knowledge of correctness of response, and processes-orientated feedback such as worked-out example answer, stating the latter is to be preferred. In the context of technology-based instruction, learners elect to view the minimum load of feedback required to achieve closure (Lefevre and Cox, 2016), yet for some question designs a correct answer may hide a multitude of flawed methods and misconceptions (Seppälä et al., 2006). Since students will not rigorously examine worked solutions when they know their answer is correct, it may be that this can only be avoided by better question design.

3.2 Identifying Misconceptions in Student Understanding

Closed question types such as multiple-choice questions can be used powerfully, provided that they're designed with a clear set of educational principles and goals in mind (Nicol, 2007). By analyzing incorrect answers of students on longer-form written exams for mathematics, (Albano and Pepkolaj, 2014) were able to develop effective recovery learning path tests on top of a fairly primitive Learning Management System quiz tool. Careful question design may be able to detect and correct student misconceptions which are understood in advance, but how can this prerequisite information for understanding be gathered? In some subjects common misconceptions amongst student bodies have already been identified through academic research initiatives, and the results compiled into documents targeted at teaching staff (one such example would be (Hestenes et al., 1992)), but what if such documents are not readily available?

Interactive questions are capable of recording a greater variety and density of information than traditional tests questions. Researchers in (Greiff et al., 2015) present a computerized assessment question where students are expected to optimize the outcome of a simulation by varying input values through graphical sliders. They show how time-series data that records actions taken during the assessment can be used to gain insight into the problem-solving process, and from this go on to draw statistical inferences about the efficacy of various solving strategies. Similarly, (Seppälä et al., 2006) presented an interactive system for examining Computer Science student beliefs about algorithms for creating a heap data-structure. The system required students to simulate the working of given algorithms by manipulating the visual representations of the corresponding data structures on a computer screen. By collecting data on the sequence of actions taken as students progressed towards this goal, it became possible to identify not just one-off slips but systematic misconceptions about the algorithms. These misconceived algorithm variants could then be simulated, and when a student's response matched the sequence generated by the simulation, advanced feedback targeted at the specific misconception could be delivered.

3.3 Self- and Peer-assessment

As class sizes increase, so does the cost of marking student work in terms of teachers' time and energy. Computer-based marking is clearly sufficient for closed, well-defined questions, but free text-questions, which are known to promote meta-cognition more effectively, pose technical challenges with respect to automated marking and feedback provision. Relevant automated systems do exist, often utilizing a vector-space model, however they may be overly harsh, fail to recognize the breadth of valid responses, mark only with respect to an example answer, and require the teacher to manually resolve ambiguities (Rodrigues and Oliveira, 2014). Progress has been made over the years to the point where such systems are used in practice, but using them to provide useful feedback to a learner is still an ongoing area of research (Van Labeke et al., 2013). Such a system also fails when the subjective qualities being measured cannot be reduced to linguistic similarity, such as in evaluations of long free-form essays or creative works. In contrast, self-assessment and peer-assessment have been shown to be effective in a variety of contexts such as

design, programming, art, and dance. Provided students are happy with the initial workload, such a strategy scales indefinitely, since the number of available reviewers grows with the number of students, and automated systems can be used in hybrid with human markers if necessary to detect unusual reviews or to provide an additional source of information which can be used to increase the quality of a final overall review (Luaces et al., 2015).

(Gehringer, 2014) examines multiple methods for improving review quality, but most involve aggregation of reviews, implying each student would be burdened with a large number of other student's answers to mark, an undesirable workload if formative assessments are to be undertaken regularly. One method of improving review quality which does not necessitate aggregation, is calibration, whereby students are presented with multiple example answers which have previously been assigned a mark by the teacher, before they go on to mark a peer's work.

Nicol and Macfarlane (Nicol and Macfarlane-Dick, 2006) make the case that even 'at-risk' students can learn to be more self-regulating, that the more learning becomes self-regulated the less students depend on external teacher support, that such students are more persistent, resourceful, confident and higher achievers, and that the quality of self-regulation can be developed by making learning processes explicit, through meta-cognitive training, through self-monitoring, and by providing opportunities to practice self-regulation.

Although anonymous peer-assessment may not be any more valid than named peer-assessment with respect to construct validity, pupils self-report lower levels of peer-pressure and feel more positive about the process (Vanderhoven et al., 2015).

3.4 Understanding Progress, Measuring Difficulty, and Assessment Design

Learning design typically starts by identifying the core aims of learning or abilities of a student to be developed based on the domain of the subject to be taught, and structuring practical activities in such a way as to develop these abilities (MacLean and Scott, 2011). Misconceptions, identified previously via empirical measures and documented (for example, (Hestenes et al., 1992)) may be explicitly addressed in order to accelerate the pace of learning. Similarly, hypothesized learning progressions that have been validated empirically may be used to scaffold the learning process more helpfully, and where appropriate, help students evaluate their own progress. These may be in the form of individual construct maps showing distinct levels of progress in the understanding of a concept, but they could also describe the relationship between individual competencies in a syllabus (Wilson, 2009). Given a dataset of student performance on topic-tagged questions over time, it is possible to mechanically extract a concept map showing the dependencies between topics, and in doing so identify conceptual bottlenecks in learning (Lin et al., 2015). (Lin et al., 2015) go on to show how such a generated map can be used to indicate a learner's progress to them, clearly showing how next to progress in terms of student learning requirements. Also of interest is the use of factor analysis to gain insight into core competency areas within a curriculum. This is hardly a new idea (Zou and Zhang, 2013), however accessible integration of it and other statistical tools of similar power into assessment systems that ensure resultant inferences are valid is a open area of research.

3.5 Spaced Repetition and Subject Mastery

The positive effects of spacing learning over time in increasing retention of knowledge are well documented and researched. Examples include; early work such as Ebbinghaus's discovery that the forgetting curve becomes shallower over time (and thus the optimal interval between learning repetitions increases with time) (Ebbinghaus, 1885); Spitzer's large scale study on Iowan school children (Spitzer, 1939); and a large body of studies and inquiry in the 1960s and 1970s including (Landauer and Bjork, 1978). Several businesses, some with equity of over 1 million USD, currently offer products based partly or wholly on a spaced repetition model – these include Duolingo, Synap, Brainscape, and Lingvist. Multiple open-source flash-card programs such as Anki and Mnemosyne exist, often based on an algorithm from the proprietary software SuperMemo, specifically SM-2, which is described in (Wozniak, 1990) and bears resemblance to the paper-based algorithm described in (Leitner, 1974).

Although the algorithm currently employed by SuperMemo at the time of writing (SM-17) is far more complicated than that of Anki (SM-2), in that it calculates intervals based on measures of stability, retrievability, and item difficulty; the only input required from the user is a self-assessment of confidence with respect to the answer; the time taken to answer the question is not used to measure information retrievability from memory.

The ability to appropriately schedule specific test questions at timely intervals, independent of larger tests themselves, is a feature which is hard to build

on top of a system which does not natively support it. For this reason, if spaced repetition functionality is desired in a system it is strongly advisable that such functionality be implemented explicitly at a fairly low level of design.

4 RECOMMENDATIONS AND REQUIREMENTS FOR A FORMATIVE SYSTEM

For subjects with well-defined domains (as defined by (Spiro et al., 1995)) it is clear that short term formative assessment and feedback can provide significant value, at the same time collecting sufficient data to adequately support various long-cycle improvement strategies without overburdening students or teachers, provided it is used appropriately. Ill-defined domains present more of a challenge for automated marking, but work may still be marked scalably through self and peer assessment. Identifying misconceptions requires comparatively rich data – at a minimum the content of student answers for detecting common incorrect responses, but ideally some sort of record of the process the student followed to solve the question. Longer timescale processes targeted at students like long-term adaptive testing and spaced repetition may to an extent be driven with data collected for and produced by short-term formative feedback. But the majority of existing solutions and research makes use of the student's own evaluation of their confidence in answering a question. As a result of the above, we hypothesize that an extensible and modular system for formative assessment which initially provides knowledge of correct response and optional elaborated feedback whilst collecting data about student answers and confidence, would be sufficient as a platform for a reasonably wide variety of long term improvement processes without over-burdening users in the short-term. The small size and traditional nature of this dataset is perhaps surprising in contrast with research interest in using alternative sources of data such as student blog frequency or conversations with peers for learning analytic purposes (e.g. (Cerezo et al., 2016)). Based on the above research, we present the following recommendations for designers of formative assessment systems.

Adoption:

1. Design for an initially low level of emotional investment and time commitment from users, producing immediate and obvious value in such a circumstance.
2. Whilst shaping institutional culture is not something a software tool can actively do, it is possible to lead towards institutional good practice by reducing relevant technical barriers in advance. This should be explicitly planned for and done.
3. Design data visualizations in such a way as to be easily explorable for non-expert users. Where trade-offs between power and accessibility are necessary, the editorial choices involved should be shaped by the processes the system is designed to support.
4. Allow users to easily export any visualizations they create or modify for inclusion in reports, etc., but unobtrusively watermark them with the name of the system used. In doing so, the analytical power of the system used will be gradually advertised throughout a faculty or institution.
5. Where teachers or other staff members are assessed via external criteria, ensure that good formative practice using the system will be evaluated positively against such criteria, by providing facilities that aid in documenting relevant evidence.

Feedback:

1. Allow teachers to provide students with "Knowledge of Correct Response"-style feedback automatically. Teachers should be able to provide a worked solution which can be shown upon request. Refraining from showing elaborated feedback unless desired avoids student fatigue.
2. Design the platform to make it easy for teachers to link feedback with specific tasks they would have to undertake in the future to perform desirably.
3. Provide statistical and visualization capabilities to express student progress through the curriculum, however that curriculum is currently conceived of.
4. Where it makes sense to do so, and is not discouraging, make it easy for teachers to frame feedback in terms of student's previous performances.

Self- and Peer- Assessment:

1. Support rich feedback types such as free text, pictures, and video. Or at least enable students to critique pictures and videos, if not produce them from within the system.
2. Allow reconfigurability within the system with respect to peer-review, especially with regards to how students are assigned work to review, and with regards to anonymisation.

Assessment Design and Process:

1. Implement hardness measures within assessments to automatically understand the relative difficulties students have with topics, and adjust teaching time per topic accordingly.
2. Where proficiencies correlate across groups of students and across time, use this information to inform learning design, ensuring sufficient time is spent on bottleneck topics for learning, and restructure course content according to any dependencies discovered empirically between topics.
3. Present teachers with reliability metrics of previously taken assessments, provided such automated statistics would not be misleading.

Spaced Repetition:

1. The platform must be capable (as an option to teachers since such a process may not always be desired) of scheduling formative assessments based on confidence that is self-reported by students in order to maximize their learning.

5 DESIGN OF AN EXTENSIBLE FORMATIVE SYSTEM

A simple illustration of our proposed architecture can be seen in Figure 2 The practicalities of question marking may vary dramatically between more advanced question types, and for this reason the responsibility for marking student answers lies not with the core back end server but with individual marking servers, one for each question type, which communicate with the core back end using standardized messages over a message broker. The client, the core back end, and the database are designed to be agnostic with respect to the types of questions supported. The database schema has a single table for all questions, with schemaless content and markscheme columns (whose structure is question type specific). All questions hold a single reference to an entry in the question types table, which holds the configuration for accessing the marking server for all questions of that type. In order for an extension author to write a new question type, first the format of data expected to be received from the student must be defined, then a marking server written that takes in this data via the message broker and returns some specific feedback (the structure of which must also be defined) and a number between 0 and 1 (this is so statistical analysis across heterogeneous question types is possible). The front-end code to produce this data must then be written, which consists of the UI of the question itself along with the business logic required to make it work. A UI component that displays the feedback specific to the question type should also be written, and it is advisable to write a custom statistical view for teaching staff to use for analyzing answers given to individual questions of that question type (the default statistical tools will no be able to understand the structure of student answers without help since that is specific to each question type). With the above completed, the relevant files can be placed in their appropriate places, a new question type created in the database with appropriate marking server configuration, and from that point on questions of that new type may be created.

Much of the desired functionality can be accomplished with the system as described above – for example, feedback is trivially possible, and collecting data about student misconceptions can be accomplished by extending the system with complex question types which can be built using web technologies. Self assessment likewise can be accomplished via a dedicated question type. However, some practices require certain outcomes to occur as the result of previous student performance and/or the time that has passed: adaptive testing assigns students recovery tests as a result of poor performance in previous tests; spaced repetition involves reassigning previously seen questions to students after an interval calculated from past confidence in their answers; peer review needs students to be assigned to review other student's work after they have submitted it. To provide sufficient functionality for these practices whilst keeping the core application simple, we will implement a conceptually separate scheduler which runs both periodically to ensure time-dependent rules are followed globally, but also upon request. Whilst teachers assigning tests and questions manually to students is handled directly by the core application, any other assignment of questions (for any reason) is the responsibility of the scheduler. The rules of the scheduler are not designed to be changed by extension authors, and will likely be fairly opaque to users; instead of seeing the exact rules that are going to be followed, teachers will be presented with a settings UI that allows them to enable, for example, *"computer adaptive testing"* at the press of a button, and this will enable all of the relevant GUI components as well as the back-end logic designed to support such a work-flow. With the above said, there is no reason a particularly dedicated researcher could not add new rules to the scheduler to support some new process.

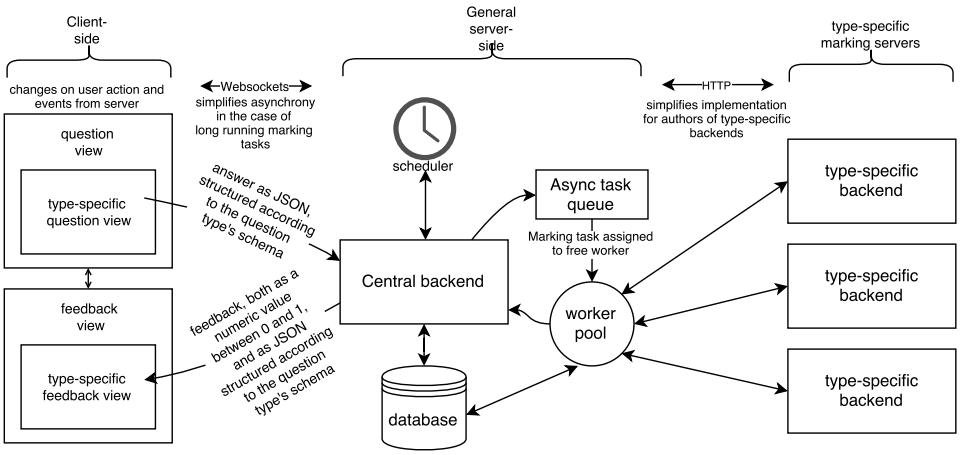

Figure 2: Design of a formative assessment system.

6 CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

By targeting a web platform, a medium ideally suited for the development of interactive and visual tools, we aim to enable learning experiences and questions to be developed with relative ease and in a way that remains coherent and integrated with the other formative practices we hope to encourage. Novel interactive questions types are an area of research with potentially rich rewards, as is applying what is known currently about the computer-assisted identification of misconceptions to new subject domains. It is our intention that our system serve as a platform for multiple researchers looking to explore the effectiveness of richly immersive web-based learning tools.

One of the largest barriers to effective use of data in an educational setting is the lack of statistical tools which are easy to use for non-experts. Because our system aims to target practices that take place over a variety of timescales whilst collecting data in a consistent format we believe it to be suited for the prototyping and subsequent empirical testing of a wide range of graphical statistical tools. Of particular note for researchers is the problem of conveying nuanced statistical ideas which affect validity of inferences made to staff who have not been statistically trained, especially doing so visually.

Appropriate treatment of issues of reliability and validity have been identified as key deficiencies in existing research ((Gikandi et al., 2011), (Bennett, 2011)). A potential deficiency with the system design presented above is the lack of resistance to questions being incorrectly labeled with the wrong topics. Despite years of research, it remains unclear how components of feedback such as timing affect its overall efficacy. It is our hope that a system such as ours once developed, will permit long-term A/B testing of variations to these components. Although ethical and professional concerns related to delivering a potentially sub-par educational experience to a section of paying students will have to be addressed in order to proceed.

REFERENCES

Albano, G. and Pepkolaj, L. (2014). Formative self-assessment to support self-driven mathematics education at university level. In *International Conference on Web-Based Learning*, pages 82–91. Springer.

Allal, L. and Schwartz, G. (1996). Quelle place pour l'évaluation formative dans l'enseignement au cycle d'orientation? *CO Infos*, 178:5–8.

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

Black and Wiliam (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2):139–144, 146–148.

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

Black, P., Harrison, C., Lee, C., Marshall, B., and William, D. (2004). Working inside the black box: Assessment for learning in the classroom. 86(1):8–21.

Blin, F. and Munro, M. (2008). Why hasn't technology disrupted academics' teaching practices? understanding resistance to change through the lens of activity theory. *Computers & Education*, 50(2):475–490.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., and Núñez, J. C. (2016). Students' lms interaction patterns

and their relationship with achievement: A case study in higher education. *Computers & Education*, 96:42–54.

Ebbinghaus, H. (1885). *Über das Gedächtnis*.

Gehringer, E. F. (2014). A survey of methods for improving review quality. In *International Conference on Web-Based Learning*, pages 92–97. Springer.

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

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? a showcase study based on the pisa 2012 assessment of problem solving. *Computers & Education*, 91:92–105.

Hall, R. and Hall, M. (2010). Scoping the pedagogic relationship between self-efficacy and web 2.0 technologies. *Learning, Media and Technology*, 35(3):255–273.

Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1):81–112.

Hestenes, D., Wells, M., Swackhamer, G., et al. (1992). Force concept inventory. *The physics teacher*, 30(3):141–158.

King, E. and Boyatt, R. (2015). Exploring factors that influence adoption of e-learning within higher education. *British Journal of Educational Technology*, 46(6):1272–1280.

Kotter, J. P. and Cohen, D. S. (2002). *The heart of change: Real-life stories of how people change their organizations*. Harvard Business Press.

Landauer, T. K. and Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. *Practical aspects of memory*, 1:625–632.

Lefevre, D. and Cox, B. (2016). Feedback in technology-based instruction: Learner preferences. *British Journal of Educational Technology*, 47(2):248–256.

Leitner, S. (1974). *So lernt man lernen*. Herder.

Lin, Y.-S., Chang, Y.-C., Liew, K.-H., and Chu, C.-P. (2015). Effects of concept map extraction and a test-based diagnostic environment on learning achievement and learners' perceptions. *British Journal of Educational Technology*.

Looney, J. (2011). Integrating formative and summative assessment: Progress toward a seamless system? *OECD Working Paper*.

Luaces, O., Alonso, A., Troncoso, A., Bahamonde, A., et al. (2015). Including content-based methods in peer-assessment of open-response questions. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 273–279. IEEE.

Macfadyen, L. P. and Dawson, S. (2012). Numbers are not enough. why e-learning analytics failed to inform an institutional strategic plan. *Educational Technology & Society*, 15(3):149–163.

MacLean, P. and Scott, B. (2011). Competencies for learning design: A review of the literature and a proposed framework. *British Journal of Educational Technology*, 42(4):557–572.

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and higher Education*, 31(1):53–64.

Nicol, D. J. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218.

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

Rogers, E. M. (1995). *Diffusion of Innovations: Modifications of a Model for Telecommunications*, pages 25–38. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

Russell, C. (2009). A systemic framework for managing e-learning adoption in campus universities: individual strategies in context. *Research in learning technology*, 17(1).

Seppälä, O., Malmi, L., and Korhonen, A. (2006). Observations on student misconceptions: A case study of the build–heap algorithm. *Computer Science Education*, 16(3):241–255.

Spiro, R. et al. (1995). Cognitive flexibility, constructivism and hypertext: Random acces instruction for advanced knowledge acquisition. *P. Steffe e J. Gale*.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, 30(9):641.

Van Labeke, N., Whitelock, D., Field, D., Pulman, S., and Richardson, J. T. (2013). Openessayist: extractive summarisation and formative assessment of free-text essays.

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., and Schellens, T. (2015). What if pupils can assess their peers anonymously? a quasi-experimental study. *Computers & Education*, 81:123–132.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment*, 11(3-4):283–289.

Wiliam, D. and Thompson, M. (2007). *Integrating assessment with learning: What will it take to make it work?* Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6):716–730.

Wozniak, P. A. (1990). Optimization of learning. *Unpublished master's thesis, Poznan University of Technology. Poznan, Poland*.

Zou, X. and Zhang, X. (2013). Effect of different score reports of web-based formative test on students' self-regulated learning. *Computers & Education*, 66:54–63.