Technology enhanced assessment in complex collaborative settings

Mary Webb
Kings College London

David Gibson
Curtin University

Abstract

Building upon discussions by the Assessment Working Group at EDUsummIT 2013, this article reviews recent developments in technology enabled assessments of collaborative problem solving in order to point out where computerised assessments are particularly useful (and where non-computerised assessments need to be retained or developed) while assuring that the purposes and designs are transparent and empowering for teachers and learners. Technology enabled assessments of higher order critical thinking in a collaborative social context can provide data about the actions, communications and products created by a learner in a designed task space. Principled assessment design is required in order for such a space to provide trustworthy evidence of learning, and the design must incorporate and take account of the engagement of the audiences for the assessment as well as vary with the purposes and contexts of the assessment. Technology enhanced assessment enables in-depth unobtrusive documentation or 'quiet assessment' of the many layers and dynamics of authentic performance and allows greater flexibility and dynamic interactions in and among the design features. Most important for assessment FOR learning, are interactive features that allow the learner to turn up or down the intensity, amount and sharpness of the information needed for self-absorption and adoption of the feedback. Most important in assessment OF learning, are features that compare the learner with external standards of performance. Most important in assessment AS learning, are features that allow multiple performances and a wide array of affordances for authentic action, communication and the production of artefacts.

1. Introduction

Our previous analysis (Webb, Gibson, & Forkosh-Baruch, 2013) following discussions at EDUsummIT 2011, identified student and teacher involvement in assessment including digitally-enhanced assessment as critical for 21st century learning. Digitally-enhanced assessments were defined as those that integrate: 1) an authentic learning experience involving digital media with 2) embedded continuous unobtrusive measures of performance, learning and knowledge, which 3) creates a highly detailed, high resolution data record which can be computationally analyzed and displayed so that 4) learners and teachers can immediately utilize the information to improve learning. This unobtrusive measuring approach is a vision of '*quiet assessment*' whose volume can be turned up by learners and teachers whenever they wish in order to check their progress.

This article, developed following further discussions of the Assessment Working Group at EDUsummIT 2013, aims to build on our previous analysis by reviewing recent developments in technology enabled assessments of collaborative problem solving in order to identify examples,

approaches and their challenges and point out where computerised assessments are particularly useful (and where non-computerised assessments need to be retained or developed) while assuring that the purposes and designs are transparent and empowering for teachers and learners.

2. Background

When the EDUsummIT Assessment Working Group met again in 2013 some of the challenges identified in 2011 remained, including uncertainty as to whether and how the following four perspectives on assessment: feedback information, improvement decisions, degree of engagement and understanding, and value judgments, can co-exist to the benefit of learners (M. Webb, E., Gibson, & Forkosh-Baruch, 2013). Even with the increased possibilities that IT provides we have not yet found a way to say confidently that the multiple purposes for which some assessments have been used (Mansell, James, & Group, 2009) can or should be supported through the same assessment systems. This is because the impacts of some purposes interact with the validation processes of others (Messick, 1994). Therefore in considering assessment design for multiple purposes, users need to examine impact factors carefully in order to minimise negative impacts on learning and learners. In this review, we assert that integration can occur to meet the multiple purposes, because the affordances of technology can redefine the nature of an assessment task, and we provide a high level outline of the processes for engaging in those considerations in the design of assessments of collaboration, particularly collaborative problem-solving, as exemplified in the Organisation for Economic Co-operation and Development (OECD) draft for the Programme for International Student Assessment (PISA) assessment of the interaction of these two domains (PISA, 2015).

Discussions at EDUsummIT 2013 led to three main recommendations. First, researchers, policy-makers and practitioners agreed to examine and promote assessment of collaborative learning in problem solving environments as an important and complex problem space both for learning and for assessment. For example, significant challenges remain for developing validation approaches that can take account of the complexity of learning experiences for collaborative group tasks. Second, we saw a need to develop theory for big data in educational research (see the article “Big data in educational assessment“ (Gibson and Webb, 2015) also in this journal’s special edition). Third, we underscored the primacy of the need to engage teachers in the design of learning analytic tools for instructional practices and in interpreting and using results. Here, we will focus on engaging teachers and students in the technology-enabled assessment of collaborative learning. In the related article, we discuss their engagement with big data.

Our reviews and group discussions of global ICT and assessment since 2009 (M. Webb, E., et al., 2013) have combined research-based findings with classroom observations of assessment practices (Black & Wiliam, 1998) and evidence-centered assessment design (ECD)(Mislevy, Steinberg, & Almond, 1999) . We examined the ECD framework because it has become quite widely used among designers of computer-based assessment as it makes explicit the interrelationships and substantive arguments among the main elements of the design and implementation: domain models, validity, assessment designs and operational processes (Mislevy et al., 2003). The framework has diagnostic capabilities and provides opportunities for stakeholders to view estimated competency levels, examine the evidence on which these judgements were based and to use this information for a variety of processes as appropriate (Shute, 2011 P.9). It is also the primary organizing theoretical framework for the PISA assessment of collaborative problem solving (Chauncey & Azevedo, 2010), which we present as an example of the principles under discussion.

The stages of the evidence-centered design process include domain analysis and modelling that defines the assessment problem space and shapes its affordances; the conceptual assessment framework that defines the assessment task, performance model and evidentiary rules that map from student performance to the domain model; the delivery and sampling plan that defines the media, range of problem space for tasks, and presentation issues.

We view a technology-enabled collaborative learning environment as a rich context for assessing higher order skills, as long as the purpose and design of the assessment is clear about its targets and the assessment tasks are constructed to include technology as part of the collaborative problem solving task and the assessment provides timely useful feedback to teachers and students.

In the sections that follow, we first review the issues concerning collaborative learning in problem-solving environments with a specific focus on science learning in compulsory education, where some important developments are taking place. Then we examine the broader contexts of technology-based assessment using a model that highlights the transformational nature of technology, with a view to considering the potential of technology-based assessments for assessing higher levels of knowledge and performance. Then we mention briefly the further challenges that will need to be addressed in order to utilise big data to assess such higher levels through quiet assessment. We expand our explanation of the significance of developments in big data research in another paper in this special issue (Gibson and Webb, 2015).

3. Collaborative learning in problem-solving environments

A focus on the assessment of collaborative problem-solving (CPS) is pertinent and timely for three main reasons. First the decision by the OECD PISA Project to assess CPS in 2015 (Chauncey & Azevedo, 2010) means that a spotlight will be on this important aspect of learning (Blatchford, Baines, Rubie-Davies, Bassett, & Chowne, 2006; Voogt, Erstad, Dede, & Mishra, 2013). PISA is a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students in order to determine the extent to which they can apply their knowledge to real-life situations and hence are prepared for full participation in society. According to PISA more than 70 economies have signed up to participate in the 2015 assessment which will focus on science, including CPS, as the major domain. Furthermore the PISA conceptual framework provides us with an example of a potentially significant step forward in computer-based assessment. Second although collaborative learning is known to have a positive impact on students' learning (Johnson, Johnson, & Stanne, 2000; Lee, Linn, Varma, & Liu, 2010) productive interactions between students are not easily achieved (Barron, 2003; Chan, 2012) and appropriate learning situations are challenging to implement (Bell, Urhahne, Schanze, & Ploetzner, 2009). Therefore CPS is a challenge for learning and teaching as well as for assessment and tackling these issues together has game changing potential for education. Third, CPS is a complex problem space that entails and entrains a great many other issues relevant to the use of IT in assessment and thus will enable us to examine further the potential for new developments in assessment.

Collaborative learning involving inquiry and problem-solving has become commonplace in curricula around the world especially in subjects such as science, maths, geography and history (Chauncey & Azevedo, 2010). However it is also generally acknowledged that collaborative learning is a challenging process for students requiring a complex set of cognitive, metacognitive and social skills in order to engage in interactive processes such as developing shared task understanding, negotiating shared perspectives, argumentation and maintaining focus (see for example Barron, 2003; Chan, 2012; Evagorou & Osborne, 2013). Studies of 11–12-year-olds working in triads have

shown that student groups often failed to achieve the productive interaction necessary for CPS (Barron, 2003). This failure was often associated with relational issues such as competitive interactions and self-focused problem-solving trajectories (ibid). In order to interact and collaborate successfully students need to self-regulate their own learning as well as being aware of the feelings and challenges of others so that they can engage in co-regulation and socially shared regulation of metacognitive, emotional and motivational aspects of learning within the group (Järvelä, Volet, & Järvenojä, 2010; Järvenoja & Järvelä, 2009; Ukan & Webb, 2014 in preparation). Recent research is beginning to enable us to understand the interactions between individual and social regulation of learning and how these affect CPS (Ukan & Webb, 2014 in preparation) but there is a need for further research to understand the relative importance of different types of regulation and how these interact across sequences of activities. These complex interactions means that managing effective CPS requires teachers to understand and develop students' individual cognitive, social and emotional capabilities, organise and structure groups in order to foster this development, devise tasks that will provide a suitable level of challenge for the group and intervene and scaffold learning in order to ensure that productive interactions are taking place. Understandably therefore, given these demanding requirements, teachers are often reluctant to utilise CPS for their students' learning due to factors such as fear of cheating or plagiarism, under-emphasis in high-status examinations, reticence by students to lower competitive advantage, the effort required to design good learning activities, and how to assess the activity (Manlove, Lazonder, & Jong, 2007).

4. Using ECD to analyse the domain and identify the problem space for assessment

The complex problem space of CPS enables consideration of the importance of the context of assessment, the role of assessment in promoting higher levels of knowledge and performance, and the role of assessment in determining what someone knows and can do. For example a question emerged in the EDUsummIT 2013 discussion, which illustrates the complexity of CPS: Is an idea substantial if it helped shape the final product by eliminating competing ideas but is not mentioned in the final outcome? This question implies the need to keep track of the *time series of the evolution of a group's process* as well as *its decisions*. Is someone's role in collaborative work completely documented in the final product? This question implies that assessment needs to *track the contribution of each person* during the process of group's evolution, not just the final group outcome. What if there is no final product; has the group not collaborated? Are we interested in both the impact of someone's collaborative skills, as well as *which skills they used during the collaboration*? The OECD has decided to pay attention to only the skills during use, not to the final result of the collaboration; but many classroom teachers are interested in the results and products created by a collaborative effort and they wonder *how to assign credit* in these situations. In formal assessments, these issues have implications for policy-makers, practitioners and researchers.

The OECD framework for constructing assessments of CPS builds upon an individual assessment of problem-solving, which was already defined and well understood in earlier PISA assessments (Chauncey & Azevedo, 2010; Sandi-Urena, Cooper, & Stevens, 2010) and conjoins that definition with a new domain framework of collaboration made operational in a simulated collaborative context. That is, to control and manipulate the variables of collaborators, a computer plays the part of collaborators while an individual displays their collaboration knowledge and skills while solving a problem shared by the simulated group. However collaboration has other contexts of interests to educators; it can include building something, co-performing as in theatre and music, changing one's mind as part of reaching a shared understanding, taking and defending sides of an issue in order to

examine an idea, and supporting other group members as they play their roles in the group's progress. So the domain model of the knowledge and skills for collaboration is potentially large. Without a computer simulating other members of a simulated group, the assessment issues can be complex, leading some instructors to avoid grading group work due to the puzzle of how to assign responsibility and ascertain individual credit (Hickey & Zuiker, 2012).

5. Using ECD to plan technology-based assessment taking account of contexts

The operational framework of any assessment has to take into account the technologies, tasks and assessment contexts in which it will be applied (Funke, 1998). We therefore purposefully use the plural term 'contexts' because both in its various external as well as internal characteristics, an assessment takes place in multiple situations (e.g. different times, classes, parts of a school, region or country) and utilizes multiple perspectives (e.g. the student, teacher, parent, board of examiners, community). Each assessment has a set of purposes linked with the methods for achieving them. For example, parents, teachers, school administrator and students all have different needs for information at different times and want to use the information for different reasons. Any assessment plan must address those external contexts while also selecting appropriate internal structures needed to elicit a valid student response or performance, score its artefact in relationship to some model of task performance, and communicate the result of the evidentiary findings in one or more contexts (Pellegrino, Chudowsky, & Glaser, 2001).

We will discuss the contexts of technology-based assessment of collaborative problem-solving in two relationships:

1. In terms of the problem space given to the student in which to perform and be assessed
2. In terms of the level of technology-in-use for the assessment

The plan for the OECD assessment of collaborative problem-solving provides an example of an expert conception of how someone solves a problem, conjoined with how they do so in a collaborative environment (PISA, 2013). This is the *problem space* of the planned assessment. To constrain the quite complex variables that would be involved if the collaboration was among a set of real people, the OECD plan is to utilize the computer to play roles as collaborators in what some would call a *virtual performance assessment* (Clarke-Midura, Code, Dede, Mayrath, & Zap, 2012).

5.1 Collaborative problem-solving contexts: the PISA 2015 model

The focus of PISA 2012 included substantial research on the development of assessment methods for individual problem solving. But there are no established methods or existing large-scale assessments of individuals solving problems in a collaborative context. So for the 2015 assessment the OECD has developed a new *domain model* (Table 1) for an assessment of individual collaboration competencies utilized during a problem-solving challenge, which draws from an established definition and methods of measuring individual problem-solving and conjoins those with three collaboration competencies described below.

The definitions shaping the domain model are:

> **Individual Problem Solving**: an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious.

> **Collaborative Problem Solving:** the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution.

The word 'agent' refers to either a human or a computer-simulated participant. In both cases, an agent has the capability of generating goals, performing actions, communicating messages, reacting to messages from other participants, sensing its environment, adapting to changing environments, and learning (Franklin & Graesser, 1996).

The domain model for the OECD assessment (Table 1) has been determined as the intersection of:

Collaboration competencies:

1. Establishing and maintaining shared understanding;
2. Taking appropriate action to solve the problem;
3. Establishing and maintaining team organisation.

Problem solving competencies:

A. Exploring and understanding
B. Representing and formulating
C. Planning and executing
D. Monitoring and reflecting

At the intersections of these two dimensions (e.g. A1, A2…D3) are specific activities that will be detected by the computer in terms of the *actions, communications,* or *products* created by the test taker; each detection will be evaluated by the evidentiary process within a finite set of levels of performance determined by and supported by the affordances of the virtual performance assessment problem space. For example, across the scenarios the test taker will face, the collaboration skills will vary across low, medium, and high difficulty levels, while the problem-solving skills will range from low to medium difficulty. It is anticipated that 5-30 measurements will be derived from each scenario. Each of these individual items will provide a score for one or more of the three CPS competency subscales.

Table 1. Matrix of Collaborative Problem Solving Skills for PISA 2015 (PISA, 2013 P. 11)

	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
(A) Exploring and Understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and Formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organisation (communication protocol/rules of engagement)
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/ being performed	(C2) Enacting plans	(C3) Following rules of engagement, (*e.g.*, prompting other team members to perform their tasks.)
(D) Monitoring and Reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organisation and roles

The contexts of various scenarios will be presented in clusters because the type of collaboration and associated rules of engagement change if the context of the collaboration is helping, working, consensus building, negotiating, debating, and participating in jigsaw configurations where group members have different information that needs to be integrated into a solution. Table 2 shows the context dimensions.

Table 2 CPS context dimensions (PISA, 2013 P.16)

Context	Dimension	States
Problem Scenario	Task type	*E.g.* Jigsaw, consensus building, negotiation
	Settings	Private vs. public Technology vs. non technology School (formal) vs. non-school (informal)
	Domain content	*E.g.* Math, science, reading, environment, community, politics
Team Composition	Size of group	2 or more (including the student)
	Symmetry of status of team members	Symmetrical vs. Asymmetrical
	Symmetry of roles: Range of actions available to each team member	Symmetrical vs. Asymmetrical
Task characteristics	Openness (c.f. PISA PS 2012)	Well-defined vs. Ill-defined
	Information availability: Does the student receive all necessary information at once? (c.f. PISA PS 2012)	Static vs. Dynamic
	Interdependency: Student A cannot solve problem without student B's acts)	Low to High
	Symmetry of goals	Group vs. individual
	Distance to solution (From beginning state to goal state)	Small, medium or large
Medium	Semantic richness	Low to High
	Referentiality to the outside world	Low to High
	Communication medium cost of grounding Interdependency: Student A cannot solve problem without student B's acts)	Low to High
	Problem space: does the student get information about other team members' actions?	Explicit vs. implicit

5.2 Implications of the OECD/PISA assessment of CPS for policy makers, teachers and learners

The OECD example promises to enable assessment of CPS skills in a controlled way thus making it possible to conduct a widespread comparative assessment across countries. In order to support interpretation of the PISA data, additional information is collected on students' backgrounds, their approaches to learning and school organisation. Typically policymakers take note of their country's performance in PISA (Davis, 2000) and are likely to review policies and practices in relation to teaching and assessment depending on outcomes of PISA tests. Policies on science education in many European countries already emphasise the importance of problem-solving, inquiry learning and collaborative engagement but actual practices are probably quite diverse (Eurydice, 2011). A number of existing instructional programmes aim to support students' acquisition and use of regulation processes during science inquiry activities (e.g., Manlove et al., 2007; Sandi-Urena et al., 2011). However, they mostly target individual aspects of metacognitive regulation, whereas research suggests (Ucan and Webb, 2014 in preparation) that it is necessary to include social, emotional and motivational aspects of regulation processes in such programmes.

While recommendations for teaching approaches emphasise collaborative learning, high-stakes assessments at the school, course and unit level focuses predominantly on assessing individuals. The importance of assessment of collaborative work is sometimes recognized, but rarely addressed, perhaps due to a bias toward a particular view of cognition and situated learning as the sole responsibility of an individual learner (Järvelä, Volet, & Järvenojä, 2010). In addition, assessments by teachers through observation, judgment, test making, and scoring, which could contribute significant information for the assessment of 21st century skills, have decreased in compulsory education because concerns about reliability and costs have outweighed those of validity, trustworthiness, and value to the learner(Harlen & Deakin Crick, 2002; Weller, 2001).

Fortunately, the PISA tasks in collaborative problem solving are a vivid example of the future of evidence-centered assessment of higher order thinking utilizing innovative ICT affordances and allowing analyses such as those outlined above. As the main PISA assessment of CPS is intended only to provide comparative data at country-level by random sampling of schools, the challenge for PISA assessments is less difficult than that of country-based assessments which try to combine assessments of individual students' progress, with comparisons between schools thus creating complex validation issues as discussed earlier.

5.3 Implications of choices made in the OECD/PISA Design for assessment of CPS for more broad-based assessments

The OECD example illustrates how the process of collaborative problem solving in a computer-based assessment can generate a complex data set that contains actions made by the team members, communication acts between the group members, and products generated by the individual and the group. Each turn can be classified into levels of proficiency for each CPS competency. Because the focus is on the individual, measurement will be on the outputs of the student, in contexts where the rest of the group provides controlled information about the state of the problem solving process and the contexts are managed to provide levels of difficulty as needed in multidimensional Rasch-modelling (Brown, 2005).

We now turn to another set of contexts that influence the construction of a technology-enhanced assessment. Consideration of this set of contexts is important if we are to apply lessons from the large-scale, tightly managed and controlled psychometric model of an assessment such as the PISA assessment of collaborative problem solving, to a broader range of formal to informal assessment practices of classrooms, school, and educational systems. To facilitate the discussion, we have

chosen a model of the integration of ICT in teaching and learning with both structural and developmental implications.

6. Technology integration contexts

The SAMR model of Reuben Puentedura (Jacob-Israel & Moorefield-Lang, 2013) describes four ways that technology can be used in teaching and learning – substitution, augmentation, modification, and redefinition. The model also describes a developmental trajectory of increasing transformation that utilizes the unique affordances of technology to accomplish new things. In the following discussion, as we traverse the four ways of using technology we will refer to three perspectives on assessment that we have discussed in previous articles (Forkosh-Baruch, Gibson, Schulz-Zander, & Webb, 2009; M. Webb, E., et al., 2013): assessment OF, FOR and AS learning. Table 3 illustrates the difference in focus between assessment FOR learning and assessment OF learning in terms of the process and results. Assessment AS learning integrates assessment into ongoing learning and has the potential to facilitate and support learning while enabling judgements of performance provided that threats to validity can be removed or alleviated (Webb et al., 2013 P.453).

Table 3. Four ways to think about assessment (Webb et al., 2013 P.453)

	PROCESS focus	RESULTS focus
Assessment FOR learning	Feedback information	Improvement Decisions
Assessment OF learning	Degree of Engagement with/understanding of process	Value Judgments

The first level in the SAMR model is 'Substitution' in which technology is used to perform the same task as before the use of computers. In an assessment OF learning for example, one could present a list of questions to be answered and multiple choice response options, just like a traditional paper and pencil test. In an assessment OF learning where a teacher's observation of a complex performance produces a score on a rubric, then the substitution level of the same task might be to have the teacher carry a mobile device and score the performance on an input page. At this level, the affordances of technology might add some efficiency, for example, it could save on paper costs.

The second level is 'Augmentation' in which the technology offers a more effective tool for doing the same task. For example, in an assessment OF learning, perhaps automated scoring of the items would make grading the tests easier for large numbers of test takers; and in the performance assessment perspective, collecting, storing and retrieving the rubric scores could be made not only more efficient, but might offer a new view on the group patterns of the scoring, helping to answer how many students passed at the highest level of the rubric. At the augmentation level, some functional benefits begin to accrue. For example, perhaps the students can privately see the teacher's rubric score immediately after it is saved and see how it compares to the anonymous accumulated scores of this performance, or in the testing example, perhaps an ongoing score on the test is revealed to the teacher, who can intervene to teach if the performance pattern indicates that most students are not performing as expected.

The third level is 'Modification' in which the technology is used to make significant functional changes to traditional practices. Note that it is NOT the technology that is making these levels appear; it is *how people envision and implement its use toward their purposes* that determines the technology level of use. In an assessment OF learning, suppose that a new purpose is introduced, of seeing someone else's answer after submitting one's own, and then in order to promote learning,

allowing the student to make any adjustment desired in a second version of the answer. The initial purpose of the item (e.g. assessment OF learning by testing the declarative memory-based knowledge of the learner) has not been violated, but now, a new data point concerning learning might be added and a shift occurs toward an assessment FOR learning. Interactions such as this, with a new and more complex assessment context surrounding each item, is much harder to do on paper, so the technology is now allowing a modification of the practice that takes advantage of technology's affordances to allow significant task redesign.

The fourth level is 'Redefinition' in which the technology allows new tasks that were previously inconceivable. In an assessment AS learning, a student might create a test item for another student and while doing so consult with an expert halfway around the world, and then present the challenge as a multimedia learning object to a peer; peers can score artifacts from anywhere at anytime and see a running aggregation of the results. In an assessment FOR learning, automated scoring of the artifact might be combined with and shaped by human scores and automated feedback might also be augmented by human feedback. In an assessment OF learning, the student does not have to answer the same number of items as all other students to be diagnosed or classified; perhaps half as many question will do, because the testing framework adapts to the learner's previous answers and goes to select the next most difficult challenge rather than a random item.

At the levels of modification and redefinition, the technological context changes from an inert to an adaptive mechanism of assessment. Many analytic challenges exist at these levels. For example, the analysis of problem based learning in a collaborative setting might involve challenges of how to segment time and events into metrics of collaboration, how to deal with causal influences that loop back to change the context of the next instant, and the problem of when to zoom into high resolution details and back out to high level aggregations of those details at different points in time (Baker, 2010; Gibson & Clarke-Midura, 2013; Rupp, Gushta, Mislevy, & Shaffer, 2010; Shaffer et al., 2009). Think about a case in which a student suggests a new idea in a collaborative group, but the group ignores that individual for most of the work time; then near the end of the time, the suggestion turns out to be the idea that rescues the group from a log-jam in solving their problem. However, the student sat for most of the group's time not contributing because her idea was ignored even though she knew it might be important. How will a collaborative assessment work here? Will its metrics of communication and evidence of group participation miss this event? Will she score high on a conceptual level but low on group participation and would an averaging methodology adequately capture what happened? What is needed to better understand this case is a time-based picture of the events as well as a relationship or influence oriented perspective.

An assessment of this kind of complex situation of collaborative problem solving has heretofore been in the province of the teacher's observational powers; and the teacher may have missed the event as well. However, at the higher levels of modification and redefinition of tasks in a highly digitized assessment environment, the event data will have been captured, so the onus is upon assessment designers (just as in traditional assessments of all kinds) to ensure that all the processes and opportunities required for a fair and adequate assessment are made available and effectively utilized. These include at a minimum, according to the evidence centered design framework (Mislevy, et al., 1999), a model of what the student is supposed to know and do, a task model that elicits and allows the student to show what they know and can do, and an evidentiary process that recognizes, classifies, and scores the evidence (Pellegrino et al., 2001). In addition we advocate a fourth principle, that the teachers and students have to be able to transparently interact with the task and performance situation and the resulting data in a way that brings them understanding of

the meaning of the assessment in the context of both its purpose and their intentions (Gibson & Webb, 2013). We address this assertion in Section 7 below.

6.1 Promoting higher levels of knowledge and performance

Higher order thinking has been discussed in the literature for some time and includes a range of thinking processes such as evaluating, analyzing and creating(Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). More recent additions to the literature on learning have added emotional processes (Goleman, 1995) and social processes such as communicating, collaboratively solving problems and critical thinking (Kay & Greenhill, 2011). A review of the cognitive science literature of the decade of the 1990's made clear that learning takes places in the intersection of a community of practice, a learner with unique characteristics, a knowledge and practice base with its own representations, language and culture, and ample timely feedback and support for metacognition (Bransford, Brown, & Cocking, 2000). Our reviews and group discussions of global ICT and assessment since 2009 have combined these research-based findings with classroom observations of assessment practices (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Black & Wiliam, 1998) and evidence-centered assessment design (ECD) (Mislevy et al., 1999) which was integral to the PISA development just outlined. Thus, collaborative problem solving is viewed as a rich context for assessing higher order skills, if the purpose and design of the assessment is clear about those targets and the assessment is constructed in a technology context that at a minimum provides significant modifications or a complete redefinition of tasks.

Our argument is that combining the performance assessment perspective (FOR learning) with ECD, as unobtrusively as possible via a quiet form of data collection without disturbing the natural actions of the learner responding to a prompt or situation, and then supporting the student and teacher in harnessing their own powers of observation and pattern-finding to validate their work, enables assessment to simultaneously address assessment FOR learning with assessment OF learning, possibly for the first time in history, allowing these competing purposes of assessment and their mechanisms to not interfere with each other. This prospect is clearly at the "Redefinition" stage of assessment technology. In the context of collaborative problem solving, higher order thinking is highly likely to be evident, the question is whether assessment task designers will know how to elicit it, recognize and classify it, and provide useful and transparent feedback to the learner and teacher concerning the evidence for this higher order thinking, and whether the technology implementation of those designs has a robust model of the student, the task and the evidence needed for the assessment.

6.2 Determining what someone knows and can do

As the OECD example illustrates, technology presents a performance opportunity and medium with affordances, scaffolds the performance with 'rescues' and path choices and quietly, and unobtrusively collects evidence of what someone knows and can do. The affordances of the assessment are crucial determinants of what someone can do and those, in turn are crucial determinants of inferences of what they know based on the evidence. This basic understanding of assessment which has been discussed in great depth in the literature, has recently been given a more transparent and operational computational framework that can hopefully re-invigorate the dialogue about the purposes and methods of assessment OF, FOR and AS learning (Gibson & Webb, 2013). Central to the new dialogue is the role of not only the technology, but the impact of having so much rich data at the disposal of designers, researchers and developers of curriculum and assessments. This leads to the need for a sea change in educational research to absorb the methods of data science while applying the game-based, scenario-oriented perspective needed to understand

the potential for virtual performance assessments. We discuss this change in the companion article in this special issue (Gibson and Webb).

7. Engaging teachers in tool design and both students and teachers in using results

We now turn to the third recommendation of the EDUsummIT working group regarding involving teachers and students in the design of tools and engaging both students and teachers in the interpretation and use of results. This involvement is important to ensure a balance of assessment purposes to include the impacts of the *contexts* of assessments (assessments 'as' learning engagements) and their *usefulness for promoting learning and performance* (assessments 'for' learning and performance improvement) in addition to their role in *determining the extent and quality* for external audiences (assessments 'of' learning). Furthermore we expect this involvement to help to avoid or mitigate some of the risks discussed later. While the approaches discussed in this paper point towards opportunities for technologies to enable assessments, OF, FOR and AS learning while students are engaged in authentic tasks, existing capabilities of computer based assessments of complex CPS skills as exemplified by the PISA 2015 approach, are still relatively limited as we have discussed. In the immediate future it is likely to be essential for assessments of these complex skills and understanding to be a shared process between technologies, teachers and learners. Therefore, in order to continue to develop and build on good practice in assessment design in classroom settings, designers of computer-based assessments need to consider not only building valid assessments but also incorporating tools that enable teachers to understand and comment on the main elements of the design. Furthermore developments in data mining, analytics and visualisation techniques are needed not only to share the outcomes of assessment but to enable "drilling down" to understand how these assessments were made. Developments in the theory of the data analyses that are required for such analytics are discussed in depth in Gibson and Webb (this special issue). Existing "learning dashboards", while they currently fall far short of being able to present the sophisticated traces at different levels of resolution that we envisage for quiet assessment of complex skills and understanding, already do provide opportunities for learners to reflect and review their learning trajectories to some extent. Furthermore these relatively limited opportunities for students to review elements of their performance have been found to improve self-assessment and increase course satisfaction (Chauncey & Azevedo, 2010). This suggests that future developments of this analytic and visualisation capability can support assessment FOR learning.

Even when we do solve some of the technical and theoretical challenges, discussed in this and our other article (Gibson and Webb), so that capabilities of computer-based assessment become more sophisticated, interactions between peers both for supporting learning and for mutual assessment and feedback are still likely to be important for developing self-assessment and student autonomy (Black, Harrison, Lee, Marshall, & Wiliam, 2002) for many if not all learners. Classroom-based research has suggested that in order to support students in developing self-assessment, peer assessment is an important precursor (ibid.). This importance of peer assessment is also linked to discussions about the nature of feedback in assessment processes and effectiveness of different types of feedback (Hattie, 2009; Wiliam, 2011). Hattie's synthesis of meta-analyses of educational interventions revealed that feedback could be one of the most powerful influences on achievement with effect sizes of 0.7 but that effect sizes across studies involving feedback were very variable and only certain types of feedback were effective. Specifically feedback was effective where it was integrated into instruction and was clear, purposeful, meaningful and linked to students'

understanding (ibid.). Likewise Wiliam (2011), in his review of feedback and assessment FOR learning argues that feedback can only be understood in the context of the overall learning situation so that feedback becomes an interactive process rather than a piece of information. Furthermore the effects of feedback can be profound but only when students are engaged in mindful activity (ibid.). One way of engaging students in this way may be Hickey & Zuiker's (2006) student-directed "feedback conversations" in which students discuss their answers and are enabled to participate in these conversations in a constructive and supportive way following modelling by the teacher. These "feedback conversations" also resemble the formative use of summative tests, one of the four key aspects of formative assessment identified in earlier research, in which students working in pairs assessed each others' responses on test items (Black, et al., 2003). Hickey & Zuiker's (2012) study identified the importance the feedback conversations of students gaining understanding of their classmates' knowledge and its limitations which not only enabled them to support each other in knowledge development but also meta-cognitively in understanding how their knowledge was developing. These kinds of interactions require a supportive classroom culture in which students feel comfortable in making mistakes and admitting their difficulties (Boekaerts & Cascallar, 2006; M. E. Webb & Jones, 2009).

Overall these approaches to student interaction discussed above represent a shift from feedback and assessment as judgements and information provided by the teacher to much more student - directed interactions albeit within a framework and scaffolding provided by the teacher or technologies. Thus we envisage that with new computer-based assessments where students can "turn up the volume" during "quiet assessment" the assessment system will encourage students to discuss the answers, examine their performance and reflect with their peers meta-cognitively on how they might improve.

Turning now to how to engage teachers in the design of assessments, there is evidence that this may present significant challenges. Findings from an in-depth longitudinal study of English and mathematics teachers in England showed that teachers' understanding of validity was very limited probably because their attention to such issues has been undermined by external test regimes, which only require them to comply and implement, rather than think about the consequential validity of the assessments (Black, Harrison, Hodgen, Marshall, & Serret, 2010). Using Crooks at al.'s (1996) chain model of threats to validity, teachers were enabled to design valid summative assessments (Black, et al., 2010). Similarly in the United States, significant improvements in consequential validity by teacher involvement in classroom level performance assessments in the 1990's has given way to acquiescence to the demands of top-down accountability in the year 2000 national policy underpinning 'No Child Left Behind' (Hickey & Zuiker, 2012). Our sister article to this one (Gibson and Webb) discusses how validity is addressed in the ECD approach. Our recommendation is to build assessment systems that enable easy examination of validity by making assessment judgements and their basis clear by designing graphical approaches to presenting the warrants that support the claims and enabling users to drill down to examine the network of beliefs and theories on which they rely.

The complex relationship of formative and summative purposes of assessment, which can overlap as the unit of analysis moves from students, to teachers, to schools and external levels of the educational hierarchy, is compounded by the varying psychometric principles needed to understand those purposes and make best use of available data from assessments. A nuanced understanding of the interplay of the purposes and associated measurement challenges is needed at all levels of the system (Hickey & Zuiker, 2012). We have argued elsewhere that students, for example, must be able to turn up or down, the volume control on how quiet (unobtrusive) or disturbing their assessment

feedback is in relationship to their intentions, readiness to utilize information, and confidence in applying lessons from the feedback to improve their performance (Gibson & Webb, 2013).

8 Risks associated with technology-based assessment in complex learning situations

So far in this article we have focused on presenting the opportunities that we expect the technological and theoretical developments in computer based assessment of complex learning situations to provide. However there are also significant risks. We discuss these with reference to automated essays scoring (AES), another important area of development of computer-based assessment but one where various automated systems are already in use (Davis, 2000). AES has been a significant area of research and development for about 15 years driven by developments in natural language processing and machine learning as well as a need to assess vast numbers of essays. Human assessment of essays is time-consuming and therefore expensive so the market for automated approaches is very lucrative (Barron, 2003). A number of commercial AES systems now exist but their use is highly controversial owing mainly to issues about their validity (Barron, 2003; Clark, Sampson, Weinberger, & Erkens, 2007; Davis, 2000). Those who oppose the use of AES such as a major organisation for writing professionals, the Conference on College Composition and Communication, have identified several key disadvantages as: 1) writing to a machine violates the essentially social nature of writing and its value as a means of human communication and this reduces the validity of the assessment; 2) since we cannot know the criteria by which computers scores the writing, we cannot know whether particular kinds of bias may have been built into the scoring and 3) if schools see writing assessment as machine-scored they will prepare their students to write for machines (Anderson, Nashon, & Thomas, 2009). Arguably all three of these concerns could apply to assessment of CPS if we do not learn the lessons from this earlier development of AES. Clearly CPS is essentially a social activity and in the implementation planned for PISA, human interaction is replaced with machine-interaction thus whether or not the constructs being assessed in this computer-based system are similar to those that might be assessed in a face-to-face situation depends on the degree to which the system is able to simulate human interaction. This is not a problem in the way in which the PISA assessment is used provided that those making use of the comparative data generated understand the constructs being assessed and their limitations. We would hope that the second of the two concerns would be addressed in our vision of technology enhanced assessment of collaborative learning through the involvement of teachers and learners in both the design of assessment and in interpretation of the assessment information provided. Design of AES systems started before ECD was elucidated (Clark, et al., 2007) and current implementations make no attempt to provide chains of reasoning for the judgements that the systems make. The third concern is potentially the most serious and is a significant problem for any high-stakes assessment that attempt to fulfil multiple purposes, as we discussed in our previous article (M. E. Webb, Gibson, & Forkosh-Baruch, 2013 in press). The issue concerns the nature of construct validity as discussed by Messick (1994) in which he explained that the validity of the constructs depends on the particular use of assessment. Consider for example if the OECD PISA assessment discussed here were to be adopted by countries to be used as high-stakes assessment in schools. In order to ensure that their students did well on the assessment, teachers might train students by having them practice their CPS by interacting with the computer rather than in real life scenarios. In this case in order for the test to have validity it would be essential for the constructs assessed through the PISA assessment to be identical to those assessed in real life problem-solving. With regard to AES it has been recognised in use of these that the constructs assessed by AES systems are not the same as those assessed by human scorers but the outcomes from the two approaches have been shown to be highly correlated (Clark, et al., 2007) and therefore the use of these assessments in these cases has been regarded by some as valid. Consider now if teachers trained their students to perform well on writing for AES systems. Since the constructs being assessed are different from those assessed by

human scorers the students are likely to become good at those skills assessed by the AES systems while neglecting skills that are only assessed by human scorers. As explained by Messick (1994) as a result of this process of adaptation by "teaching to the test" the two approaches will gradually become less highly correlated and the AES system will no longer be valid.
This discussion of controversy over AES illustrates the potential problems for the development of technology enhanced assessment of collaborative learning if these issues are not understood by policymakers and if commercial considerations are allowed to dominate. However the vision that we have outlined in this article does, we believe, provide for a way of exploiting the opportunities provided by technological developments while mitigating the risks and at the same time supporting learning as well as assessment.

9. Conclusion

In this article we have discussed recent developments in the assessment of complex knowledge and understanding through the analysis of the OECD PISA design for the assessment of collaborative problem solving in 2015. Our analysis shows that CPS requires a complex task setting with both higher order thinking and social relationships combining to impact learning. The OECD PISA design is a significant step forward in enabling comparative assessment of CPS skills across countries. As we have discussed, this assessment is achieved by simplifying the complexity of collaborative interaction through the use of simulated behaviour of groups and individuals to enable a controlled assessment of individual's CPS skills. Analysis, using ECD, of the domain, problem space and contexts of the PISA 2015 model together with a discussion of the application of the SAMR model of technology integration to assessment OF, FOR and AS learning in the context of CPS has provided a high level outline of the processes and challenges that would be involved in enabling *quiet assessment* of CPS in an authentic context. As we have seen, the main benefits of technology enhanced assessments would be achieved at the levels of modification and redefinition of the SAMR model where opportunities for combining assessment AS, FOR and OF learning exist. Technical and analytical challenges at these levels include how to segment time and events into metrics of collaboration, how to deal with causal influences with feedback effects and when to zoom in to high-resolution details in order to identify and characterise significant contributions from group members.
In order to achieve the learning benefits that should accrue from these quiet assessments teachers and students will need to be engaged not only in the production of new tools that visualize the information (e.g. to help shape how the new tools provide the most useful and understandable information), but also in the dynamic creation of meaning from the use of those tools in learning situations (e.g. to create personal insights from the experiences as well as the reflections made possible by the new tools). This implies that teachers will need to develop their assessment literacy but these tools in themselves can and should be designed to support teachers in this development. Furthermore, reconceptualising assessment design as a shared process involving teachers enables a focus on the purposes of assessment, with due consideration of validity, while at the same time considering the optimum ways of combining technology enhanced assessment with other methods in order to achieve those purposes. Thus our vision is for a future in which technology supports teachers and students working together with technologies to understand their learning needs, move their learning forward and develop evidence of their achievements.

10. References

Anderson, D., Nashon, S., & Thomas, G. (2009). Evolution of Research Methods for Probing and Understanding Metacognition. *Research in Science Education, 39*(2), 181-195.

Baker, R. S. J. (2010). Data Mining for Education. *International Encyclopedia of Education, 3*, 112-118.

Barron, B. (2003). When Smart Groups Fail. *Journal of the Learning Sciences, 12*(3), 307-359.

Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2009). Collaborative Inquiry Learning: Models, tools, and challenges. *International Journal of Science Education, 32*(3), 349-377.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215-232.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the Black Box: Assessment for Learning in the Classroom*. London: King's College, London, Department of Education & Professional Studies.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: putting it into practice*. Buckingham, UK: Open University.

Black, P., & Wiliam, D. (1998). *Inside the Black Box: Raising Standards Through Classroom Assessment*: King's College.

Blatchford, P., Baines, E., Rubie-Davies, C., Bassett, P., & Chowne, A. (2006). The effect of a new approach to group work on pupil-pupil and teacher-pupil interactions. *Journal of Educational Psychology, 98*(4), 750-765.

Bloom, B. S., Englehart, M. B., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives, the classification of educational goals - Handbook I: Cognitive Domain*. New York: McKay.

Boekaerts, M., & Cascallar, E. (2006). How Far Have We Moved Toward the Integration of Theory and Practice in Self-Regulation? *Educational Psychology Review, 18*(3), 199-210.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience and school*. Washington: DC: National Academy Press.

Brown, N. J. S. (2005). *The Multidimensional Measure of Conceptual Complexity*. Berkeley, CA: Bear Centre.

Chan, C. K. (2012). Co-regulation of learning in computer-supported collaborative learning environments: a discussion. *Metacognition and Learning, 7*(1), 63-73.

Chauncey, A., & Azevedo, R. (2010). Emotions and Motivation on Performance during Multimedia Learning: How Do I Feel and Why Do I Care? In V. Aleven, J. Kay & J. Mostow (Eds.), *Intelligent Tutoring Systems* (Vol. 6094, pp. 369-378): Springer Berlin Heidelberg.

Clark, D., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic Frameworks for Assessing Dialogic Argumentation in Online Learning Environments. *Educational Psychology Review, 19*(3), 343-374.

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M., & Zap, N. (2012). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*, 125-148.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3*(3), 265-285.

Davis, E. A. (2000). Scaffolding students' knowledge integration: prompts for reflection in KIE. *International Journal of Science Education, 22*(8), 819-837.

Eurydice, A. (2011). *Science Education in Europe: National Policies, Practices and Research*.

Evagorou, M., & Osborne, J. (2013). Exploring young students' collaborative argumentation within a socioscientific issue. *Journal of Research in Science Teaching, 50*(2), 209-237.

Forkosh-Baruch, A., Gibson, D., Schulz-Zander, R., & Webb, M. (2009). *ICT in Teaching and Learning*. The Hague, NL: EDUSUMMIT 2009.

Funke, J. (1998). Computer-based Testing and Training with Scenarios from Complex Problem-solving Research: Advantages and Disadvantages. *International Journal of Selection and Assessment, 6*(2), 90-96.

Gibson, D., & Clarke-Midura, J. (2013). Some Psychometric and Design Implications of Game-Based Learning Analytics. In D. Ifenthaler, J. Spector, P. Isaias & D. Sampson (Eds.), *E-Learning Systems, Environments and Approaches: Theory and Implementation*. London: Springer.

Goleman, D. (1995). *Emotional Intelligence*. New York: Bantam Dell.

Harlen, W., & Deakin Crick, R. (2002). *A systematic review of the impact of summative assessment and tests on students' motivation for learning*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Hattie, J. A. C. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement* Abingdon: Routledge.

Jacob-Israel, M., & Moorefield-Lang. (2013). Redefining technology in libraries and schools : AASL Best Apps, Best Websites, and the SAMR Model. *Teacher Librarian, 42*(2), 16-19.

Järvelä, S., Volet, S., & Järvenojä, H. (2010). Research on Motivation in Collaborative Learning: Moving Beyond the Cognitive–Situative Divide and Combining Individual and Social Processes. *Educational psychologist, 45*(1), 15-27.

Järvenoja, H., & Järvelä, S. (2009). Emotion control in collaborative learning situations: Do students regulate emotions evoked by social challenges. *British Journal of Educational Psychology, 79*(3), 463-481.

Johnson, D. W., Johnson, R. T., & Stanne, M. B. (2000). *Co-operative Learning Methods: A Meta-Analysis*. Minneapolis: University of Minnesota

Kay, K., & Greenhill, V. (2011). Twenty-first century students need 21st century skills *Bringing schools into the 21st century* (pp. 41-65): Springer.

Lee, H.-S., Linn, M. C., Varma, K., & Liu, O. L. (2010). How do technology-enhanced inquiry science units impact classroom learning? *Journal of Research in Science Teaching, 47*(1), 71-90.

Manlove, S., Lazonder, A., & Jong, T. (2007). Software scaffolds to promote regulation during scientific inquiry learning. *Metacognition and Learning, 2*(2-3), 141-155.

Mansell, W., James, M., & Group, t. A. R. (2009). *Assessment in schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme*. London: Economic and Social Research Council:Teaching and Learning Research Programme.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J., Steinberg, L., & Almond, R. G. (1999). Evidence-centered assessment design. Retrieved from http://www.education.umd.edu/EDMS/mislevy/papers/ECD_overview.html

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: Committee on the Foundations of Assessment, Board on Testing and Assessment, Center for Education, National Research Council.

PISA. (2013). *PISA 2015 DRAFT COLLABORATIVE PROBLEM SOLVING FRAMEWORK*: Organisation for Economic Co-operation and Development (OECD).

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning, and Assessment Volume, 8*(4).

Sandi-Urena, S., Cooper, M. M., & Stevens, R. H. (2010). Enhancement of Metacognition Use and Awareness by Means of a Collaborative Intervention. *International Journal of Science Education, 33*(3), 323-340.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., et al. (2009). Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning. *International Journal of Learning and Media, 1*(2), 33-53.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Ukan, S., & Webb, M. E. (2014 in preparation). Social regulation of learning during collaborative inquiry learning in science: How does it emerge and what are its functions?

Voogt, J., Erstad, O., Dede, C., & Mishra, P. (2013). Challenges to learning and schooling in the digital networked world of the 21st century. *Journal of Computer Assisted Learning, 29*(5), 403-413.

Webb, M., E., Gibson, D., & Forkosh-Baruch, A. (2013). Challenges for information technology supporting educational assessment. *Journal of Computer Assisted Learning, 29*(5), 451-462.

Webb, M. E., Gibson, D., & Forkosh-Baruch, A. (2013 in press). Challenges for Information and Communications Technology supporting Educational Assessment. *Journal of Computer Assisted Learning*.

Webb, M. E., & Jones, J. (2009). Exploring tensions in developing assessment for learning. *Assessment in Education: Principles, Policy & Practice, 16* (2), 165-184.

Weller, J. (2001). Building validity and reliability into classroom tests. *NASSP Bulletin, 85* (622), 32-37.

Wiliam, D. (2011). What is assessment for learning? *Studies In Educational Evaluation, 37*(1), 3-14.