Assessment in interactive learning environments

Michael Timms

Australian Council for Educational Research

Dr Michael Timms directs the Assessment and Psychometric Research Division at ACER, which develops high-quality assessments and conducts cutting-edge research in educational measurement. He is Chief Investigator in the Science of Learning Research Centre, in which ACER is a lead institution. He is a recognised leader in the development of innovative ways to assess students in electronic-learning environments. His research is widely published in peer-reviewed journals and he was awarded the 2013 Journal for Research on Science Teaching Award by the National Association for Research on Science Teaching.

Prior to joining ACER, Dr Timms was Associate Director of the Science, Technology, Engineering and Mathematics (STEM) Program at WestEd, a pre-eminent educational research and development organisation in the United States. He led large-scale research studies in STEM education, with special focus on computer-based assessment projects, especially through the SimScientists research program (www.simscientists.org). He has been involved in the development of two assessment frameworks for the US National Assessment of Educational Progress for which he was awarded the Paul Hood award for excellence in educational research at WestEd.

Dr Timms has experience in leading evaluation research projects for other educational research grant recipients, such as universities, and has managed large-scale item development projects across many content areas. He is knowledgeable about the education systems of Australia, the United States and the United Kingdom.

Dr Timms is an Associate Editor for the *Australian Journal of Education.*

Jason Lodge

The University of Melbourne

Dr Jason Lodge is a psychological scientist, research fellow in learning sciences and an experienced lecturer in psychology and higher education. Jason is part of the Melbourne Centre for the Study of Higher Education and is affiliated with the national Australian Research Council funded Science of Learning Research Centre at the University of Melbourne. He was also the recipient of the 2014 Australian Psychological Society Early Career Teaching Award. Jason's areas of expertise are in cognitive science and educational psychology. His research focuses on the cognitive and emotional factors that influence student learning and the student experience in educational settings, particularly in digital learning environments.

Abstract

There is an increasing interest in using digital technologies to create interactive learning environments (ILEs) that both teach *and* assess student skills that are hard or impossible to assess using 'static' items such as traditional, multiple-choice questions. These interactive learning environments try to do two things simultaneously: firstly, to monitor the learning of the student in real time, providing feedback to help the student progress through the learning task; and secondly, to use the information gathered during the learning to make judgements about where the student is in learning of the topic. Essentially, ILEs draw upon the same source of data — the interactions of the student with the learning materials and embedded assessment tasks — to perform these measurements. To make these kinds of decisions, ILEs collect and analyse many variables; the complexity of these data demands the use of sophisticated assessment methods that differ from those used in traditional paper-and-pencil tests. The complexity of the ILEs also introduces challenges such as students becoming confused or failing to comprehend the feedback from the system.

Through reference to examples of ILEs, this session shows how assessment of learning takes place, how such assessment can provide valid and reliable measures, what we are learning about students' use of the systems and how we are working to refine the systems of the future.

Much of the work in the design and implementation of interactive learning environments (ILEs) with embedded assessments has occurred in science education. The reason for this is that science education worldwide has increasingly focused on ensuring that students acquire not only the knowledge and conceptual understanding of the discipline, but the practices of science that follow the scientific method. Science practices typically include the application of such skills in the early grades as recognising patterns and formulating answers to questions about the world. As they move on through the grades, students are expected to be able to gather, describe and use information about the natural world, and eventually to design experiments. This is being achieved through the use of digital materials that provide active and interactive learning scenarios in which students can apply what they have learned and engage in these science practices.

The United States has been particularly active in this area. The publication in 2012 of *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* and subsequent publication in 2013 of the *Next Generation Science Standards: For States, By States* called for a change in the way science is taught and assessed in the US. The framework advocated for a system of kindergarten to Year 12 science education that reflects the way that scientists work and think. It also called for research-based instruction that leads students to build conceptual understandings in science as they progress through their education. The framework emphasised an interweaving of the practices, crosscutting concepts and core ideas into the curriculum, instruction and assessment of the various disciplines of science. It used the term 'three-dimensional science learning' to refer to the integration of these dimensions. This three-dimensional science learning approach to science education also forms the basis of the *Next Generation Science Standards* (NGSS), which set out performance expectations that specify goals about what students should know and be able to do at each grade level.

To assist those who wish to design assessments of the NGSS, in 2014 the Committee on Developing Assessments of Science Proficiency in K–12 published *Developing Assessments for the Next Generation Science Standards* (National Research Council, 2014). The report refers to the need for classroom-based assessments that can form part of the overall assessment systems for science and this has led to many research projects in the US that have developed prototype systems.

In this paper, we show an example from the US of an ILE that both teaches and assesses simultaneously, and illustrate the kinds of measurement methods that are used to assess the learning that takes place. We also examine whether the assessment that takes place in this ILE can provide reliable measures. Finally, we discuss what has been learned to date about students' use of such ILEs with embedded assessments and the implications for design of future systems.

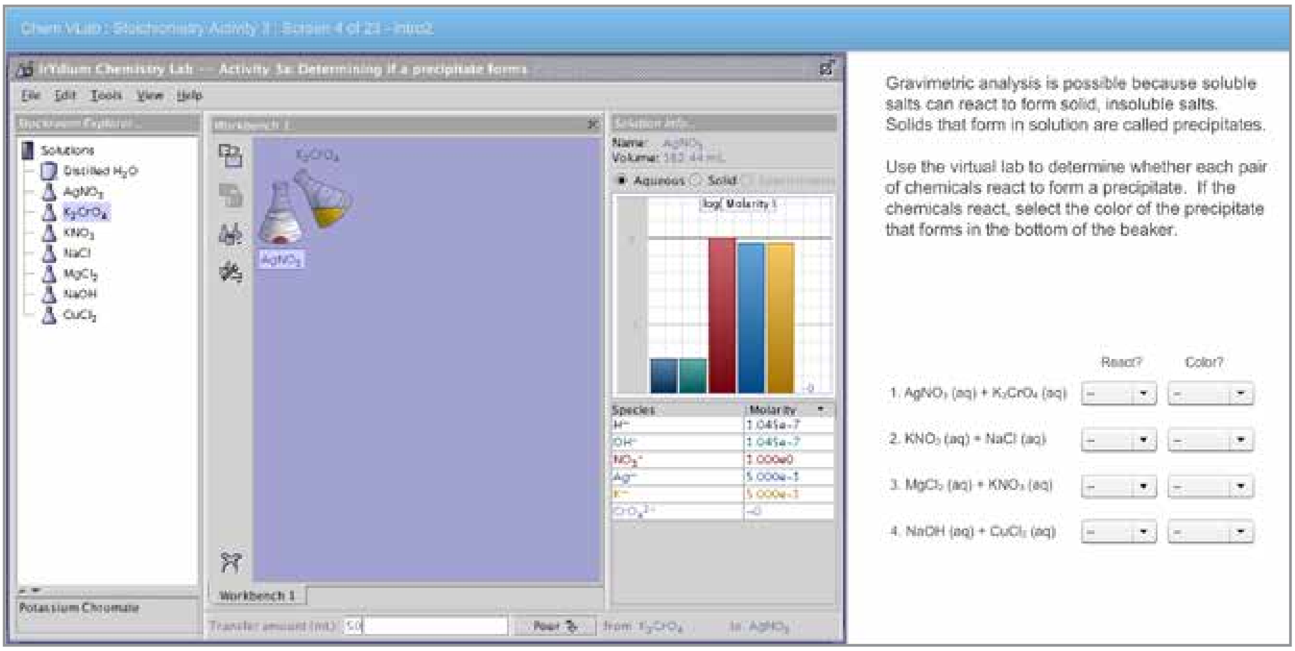
An example of an intelligent learning environment with embedded assessments

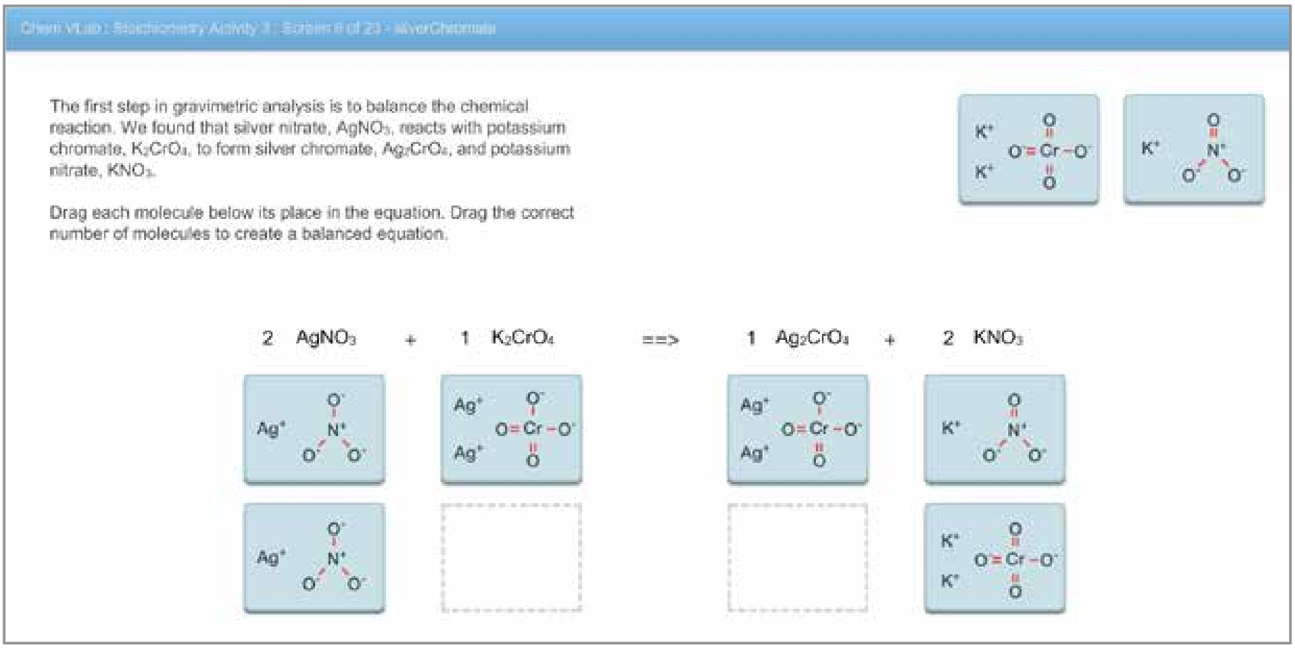
A genre of ILEs that has emerged is science learning modules based upon simulations of natural phenomena. Simulations have been chosen for science instruction because they offer some advantages. They can provide dynamic representations of spatial, temporal and causal phenomena in science systems. They can show things that are not directly observable, such as erosion over time, and they allow learners to explore and manipulate scenarios. Simulations also have the advantage of being able to present content in multiple representational forms, which has been shown in numerous studies to help students to build mental models of concepts and principles. In addition to having advantages for student learning, simulations offer advantages for assessment too. They offer the opportunity to design assessments of systems thinking, model-based reasoning and scientific inquiry which are seldom tapped in static, conventional tests. In other words, simulations offer opportunities to examine the learning process in addition to learning outcomes.

Another use of simulations in science is to provide virtual laboratory equipment that mimics what a student may find in a real science lab. The ChemVLab+ project (www.chemvlab.org), for example, provides chemistry activities that encourage students to solve authentic problems by designing experiments in a virtual chemistry lab (Davenport, Powers & Rafferty, 2014). Figure 1 shows two screenshots from an activity in the stoichiometry module. The top screenshot shows, on the right, the questions that students have to answer and, on the left, the virtual laboratory workbench in which they can select glassware, equipment and chemicals to conduct the procedures necessary to answer the questions. Rather than replacing classroom lab experiences, the ChemVLab+ activities are designed to replace lectures and traditional paper-and-pencil exercises. In the bottom screenshot of Figure 1, students are able to drag tiles that represent molecules to create a balanced chemical equation, a task that is not easy to do in paper-and-pencil tasks.

Each of the four activities in ChemVLab+ use a constraint-based modelling approach in which the errors that a student makes provide information about what the student knows and the kind of help the student needs. The data for these decisions are gathered from the student's interactions with the activities and initiated

Figure 1 Screenshot of an activity from the stoichiometry module of the ChemVLab+ project. Top, students combine chemicals in the virtual lab to determine how the chemicals react. Bottom, students drag molecules to create a balanced chemical reaction. (http://chemvlab.org)





when the student clicks the 'hint' button or attempts to move on with incorrect responses. The learner receives tiered feedback in three levels. The student is first shown where errors have been made. Next, the student is told what scientific principles are relevant to the given problem. If the student continues to make errors, the hints provide the correct response with an explanation. Student proficiency is estimated using the number of errors they make on the concepts and skills that are the targets of instruction for the module. When a class has completed the activity, teachers can access reports that indicated areas of mastery and difficulty for students. See Figure 2 for an example of the summary report that teachers receive.

The question arises as to how reliable an assessment that is embedded in a complex learning environment can be. To test this, student response data from 1373 students from eleven US high schools that used the stoichiometry module has been modelled using item response modelling. The schools were a mix of urban, suburban and rural with a range of students from low to high socioeconomic status. Item response modelling is a method used to produce estimates of student ability in a wide range of assessments including large-scale assessments like Australia's National Assessment Program — Literacy and Numeracy, for example. The data included dichotomous data points from across the four activities in the unit and scores from across the written responses in the four activities, which were scored by humans using rubrics. There were ten written response items: two items were scored 0, 1, 2 and eight items were scored 0, 1.

First, a unidimensional model that represented the whole of stoichiometry was applied to the dichotomous items and to the combined dichotomous and written response items. Two items (one dichotomous and one written response) that had psychometric characteristics outside the acceptable range were omitted from the analyses. The reliability (EAP) for the dichotomous items on their own was 0.93, and with the inclusion of the human-scored written items, the reliability increased to 0.95, a high level of reliability. A multidimensional analysis that produced student ability estimates for each of the seven content dimensions of stoichiometry was also conducted. The reliability estimates for each sub dimension are also good, demonstrating that the reports to teachers on what students know in these content dimensions are reliable to act upon. The reliability estimates are shown in Table 1.

Figure 2 Example of summary report for teachers (http://chemvlab.org)

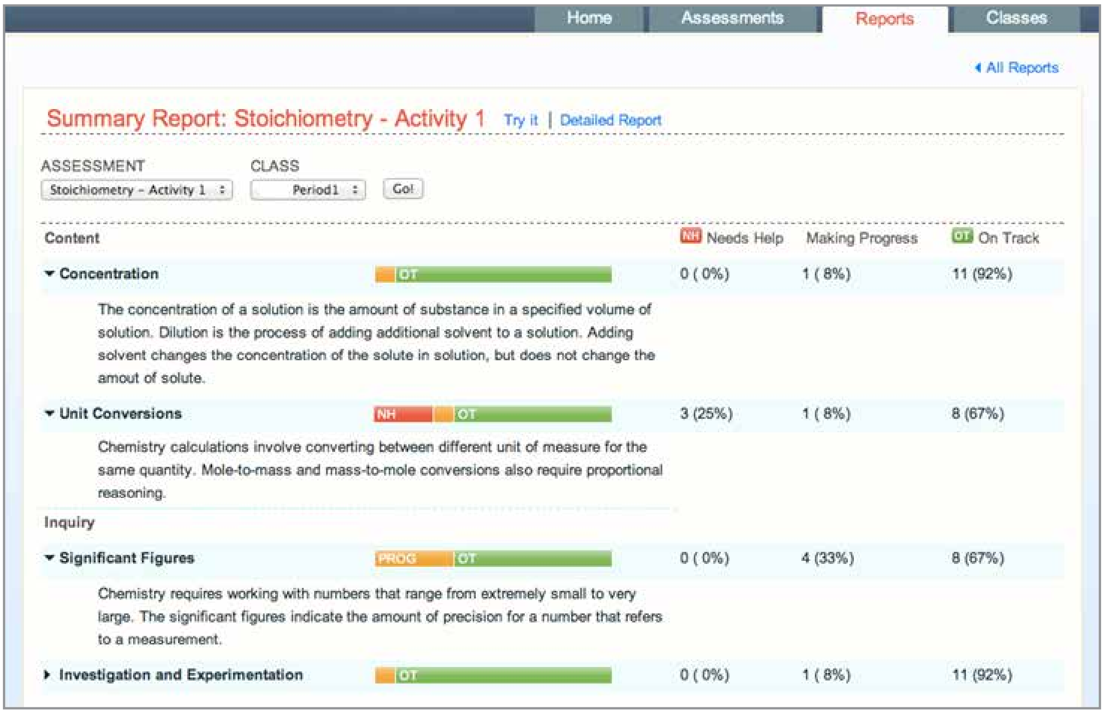


Table 1 Stoichiometry content dimensions

Dimension	# Items	EAP reliability
Concentration	20	0.85
Unit conversion	34	0.92
Molar mass	22	0.84
Balanced reactions	22	0.87
Using stoichiometry	11	0.81
Significant figures	14	0.87
Experimentation	31	0.86

Things to consider in designing embedded assessment systems in ILEs

To investigate the use of interactive assessments like those embedded in ILEs, De Boer et al. (2014) conducted a comparison of three modes of assessment for middle school students studying ecosystems. The study examined the comparative effectiveness of assessment tasks and test items presented in online modules that used either a static, active or interactive modality. A total of 1836 students used the assessments as part of normal classroom activities, taking assessments in the three different modalities on three consecutive days. The assessments tested key concepts about ecosystems and students' ability to use inquiry skills in an ecosystems context. Figure 3 shows a comparison of the three types of static, active and interactive items and how they can be targeted to assess the same learning goals. The modalities varied in how much activity students saw on the screen and how much interaction and control students had in the testing environment. Also, the interactive modality allowed for some items in which the students were given the opportunity to apply their knowledge of the targeted learning goal by, for example, designing and running their own experiments. The equivalent item in the static and active modalities only asked students to evaluate and select correctly designed experiments.

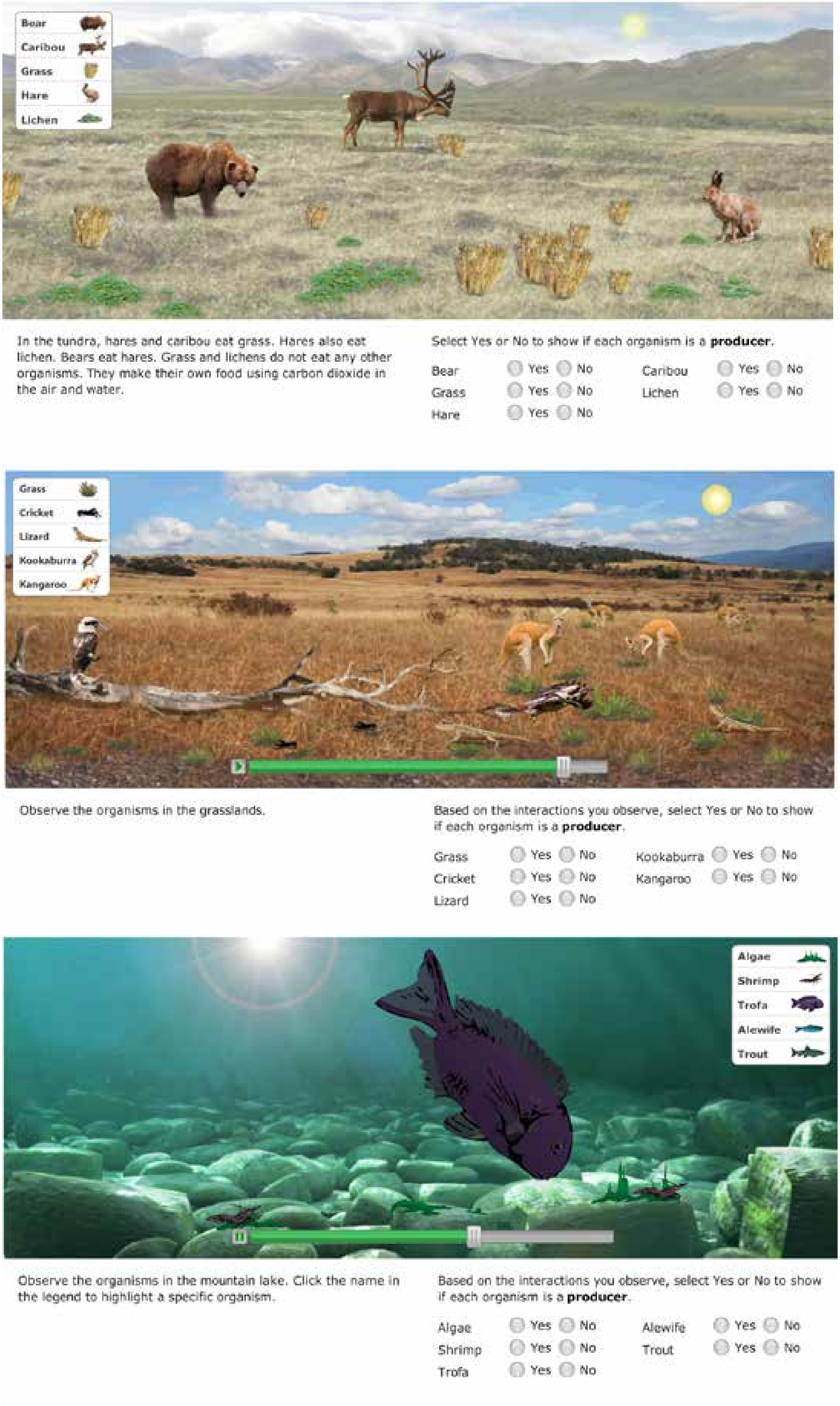
De Boer et al. (2014) found that there were no significant differences in performance on two essentially identical items that appeared in all three modalities. However, in two different sets of items on which there were differences in the activity/interactivity of the items, students performed better on the static items than on the active and interactive modality items.

De Boer et al. suggest that there are two possible explanations: that the students had more difficulty with the content of the active and interactive items, or that they had difficulty with the technology. If content is the reason, then the interactive test may be tapping into more cognitively complex skills (for example, carrying out experiments compared to identifying a correct design). Alternatively, the active and interactive items may also require a higher degree of technical experience with interactive systems. In observations of some of the students using the interactive system, De Boer et al. noted that students did not always use the technology in the way it was intended that they should.

A number of students, for example, did not immediately understand how one feature that allowed them to inspect the graphs of results worked. Also, students did not go back to rerun simulations of the ecosystems but preferred to trust their memories of what they had just seen. This points to differences in the way that students may interact with the systems in which the assessments are embedded. This may be related to observations in other research on interactive learning environments.

By the time students are into their middle years of schooling, they have had much exposure to selected response assessment items, such as multiple-choice, in which they have to evaluate some choices and select the best answer. There is no level of confusion in such items, other than that caused by the content. This is not so as we move into complex interactions in ILEs where design decisions have been made about how a simulation may work within the limitations of the screen size and the interactions possible through a keyboard and mouse or a touch screen.

Figure 3 Comparison of an item set in three modalities. Items differ in activity and interactivity (De Boer et al., 2014).



One main advantage simulations in particular offer is insight into the way that students approach and work through different content. For example, Dalgarno, Kennedy and Bennett (2014) found that, when given a simulation on blood alcohol concentration, higher education students tended to either take a highly systematic approach or a haphazard and unsystematic approach to working through the simulation. The students who relied on a systematic approach to understand the material performed significantly better in post-tests than the unsystematic group. In this instance, there was a distinct advantage to taking a scientific and systematic approach to understanding the material that was reflected in the behaviour students demonstrated in the simulation. This behaviour was also evident in the data captured during their learning and could thus be assessed.

While a systematic approach is useful in understanding many scientific concepts, in other cases students need to develop insight about a concept that necessitates a different way of thinking about it. Counterintuitive concepts such as Newton's second law provide one example of this issue. Students often need to go through some form of cognitive disequilibrium or confusion before they can reconcile the new, counterintuitive information and their intuitive experience of the world to achieve conceptual change. In a similar vein to differences in approach found by Dalgarno and colleagues, evidence that students are experiencing this confusion and achieving conceptual change can be collected and examined in ILEs (D'Mello et al., 2014). Therefore not only can the conceptual change process be monitored and assessed in ILEs, personalised feedback can be given to students at the exact point at which they need it.

In systems that use feedback we also see differences in how students use the available help and how they process it. For example, recognising the need for help is a metacognitive skill that requires students to monitor their own progress and understanding (Aleven & Koedinger, 2000). Student ability also is a factor that influences how students perform in ILEs. There is research to suggest that higher-ability learners do better within computer-mediated environments that allow for more learner control, compared to lower-ability students who do not (Recker & Pirolli, 1992). Also, those students with higher ability have been shown to be better at using help after errors, compared to their lower-ability peers (Wood & Wood, 1999). Mason and Bruning (2001) showed that students with low achievement levels perform better on both simple and complex tasks when feedback is immediate. However, students with high achievement levels perform better with delayed feedback, particularly on complex tasks. So, as we transition to interactive learning environments with embedded assessments that offer feedback, there are more design considerations to be made than in traditional assessments.

Conclusions

Interactive learning environments allow learners to engage in tasks that are able to simulate aspects of real-life scenarios and have consequently been used in a variety of science learning materials. They have been found to be useful in representing science phenomena that may be hard to observe in the classroom, such as an ecosystem, or to allow rapid and safe use of virtual laboratory equipment to conduct simulated experiments. Progress has been made in embedding assessment tasks into these learning environments which make use not only of students' responses to traditional tasks such as selecting a correct response or typing in an answer, but also in monitoring their interaction with the components of the system. Embedded assessments that occur in real time can be evaluated immediately by the learning system and therefore can offer feedback to the learner, creating a strong formative assessment. They have also been used to provide summative feedback to the learner about their overall progress and to the teacher about the progress of the class as a whole or groups within the class. The assessments have also been shown to have acceptable psychometric qualities that confirm that they can produce reliable measures and that sound judgements can be made about learners using these methods. While progress has been made, it is still relatively early days for such interactive assessments and we are still learning that there are design choices in creating such assessments so that learners can derive learning benefits from them. Finally, we know that interactive assessments take a lot more time and effort to develop, and so we need to ensure that we use them for assessment of learning in areas that are hard or impossible to assess with active or static items.

References

Aleven, V. & Koedinger, K.R. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson & K. VanLehn (Eds.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems* (pp. 292–303). Berlin: Springer Verlag.

Dalgarno, B., Kennedy, G. & Bennett, S. (2014). The impact of students' exploration strategies on discovery learning using computer-based simulations. *Educational Media International*, 51(4), 310–329. doi: 10.1080/09523987.2014.977009

Davenport, J.L., Powers, J. & Rafferty, A. (2014, April). *Classroom factors and students learning with online virtual chemistry lab activities*. Paper presented at American Educational Research Association conference, Philadelphia, PA.

De Boer, G.E., Quellmalz, E.S., Davenport, J.L., Timms, M.J., Herrmann-Abell, C.F., Buckley, B.C., Jordan, K.A., Huang, C-W. & Flanagan, J.C. (2014). Comparing three online testing modalities: Using static, active, and interactive online testing modalities to assess middle school students' understanding of fundamental ideas and use of inquiry skills related to ecosystems. *Journal of Research in Science Teaching*, *51*(4), 523–554.

D'Mello, S., Lehman, B., Pekrun, R. & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29, 153–170. doi: 10.1016/j.learninstruc.2012.05.003

Mason, B. & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us*. http://dwb4.unl.edu/dwb/Research/MB/MasonBruning.html

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Committee on Developing Assessments of Science Proficiency in K–12. Board on Testing and Assessment and Board on Science Education, J.W. Pellegrino, M.R. Wilson, J.A. Koenig & A.S. Beatty, (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Next Generation Science Standards (NGSS) Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: The National Academies Press.

Recker, M. & Pirolli, P. (1992). Student strategies for learning programming from a computational environment. In C. Frasson, G. Gauthier & G.I. McCalla (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 382–394). Berlin: Springer Verlag.

Wood, H. & Wood, D. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, 33(2), 153–169.