Enhancing the role of assessment in curriculum reform in chemistry

Thomas Holme[a], Stacey Lowery Bretz[b], Melanie Cooper[c], Jennifer Lewis[d], Pamela Paek[e], Norbert Pienta[f], Angelica Stacy[g], Ron Stevens[h] and Marcy Towns[i]

Received 10th November 2009, Accepted 4th January 2010
DOI: 10.1039/C005352J

The role of assessment in the chemistry classroom is ultimately tied to the nature of the assessments available for use. Because they provide data that can inform decisions about curricular changes, or new pedagogies, the incorporation of new assessment strategies can play an important role in how educational and curriculum reform is carried out. Several recent developments in assessment have been tied together to investigate the benefits of using multiple assessment strategies in decision making about teaching innovation. These new tools include measures of student problem solving, metacognition, cognitive development within the chemistry content at the college level and evaluation of students in affective aspects of learning. Summaries of how these new tools may be combined and what measures arise from such combinations are presented.

Keywords: assessment, problem-solving, metacognition, research-based curriculum reform

Introduction

Changes in pedagogy usually occur when an individual instructor tries a new idea that seems likely to help students learn. This manner of educational innovation is natural in so far as teaching is a fundamentally personal activity. Teachers infuse their efforts with their own personality, and thus any change in teaching is ultimately tied in some way to that personality. One common personality trait of scientists, namely a tendency to be analytical and data driven, does not always translate into classroom decisions (Cooper, 2007).

One barrier to data driven decision-making in curricular reform lies in the disjointed development of assessment tools (Labov, 2007). While there are some examples of well established and widely known assessments, such as those produced by the American Chemical Society Exams Institute[1], it is arguable that most chemistry faculty are aware of only a limited number of assessment strategies or assessment instruments (Towns, 2010). Chemistry presents a particularly good topic for improved assessment because it is (a) a component of the curriculum in a large number of science-based majors and; (b) a field with a strong mix of both qualitative and quantitative concepts.

This circumstance should not suggest, however, that advances in assessment within chemistry have not occurred. A number of important developments have been reported in the past several years, and efforts to tie these projects together are now taking shape. Examples of recently developed or characterized instruments include the measurement of student problem solving strategies and interventions to improve them using the IMMEX system (Stevens and Palacio-Cayetano, 2003; Stevens *et al.*, 2004; Soller and Stevens, 2007); the ChemQuery project for assessment tied to cognitive development of students as well as their content knowledge (Claesgens *et al.*, 2008); the CHEMX instrument to measure student expectations of the learning environment of chemistry, particularly as it compares with faculty expectations for learning (Grove and Bretz, 2007; Mathew *et al.*, 2008); an instrument similar to CHEMX called C-LASS (Barbera *et al.*, 2008); an instrument to measure student metacognitive awareness and the implications of that awareness in strategies students use for problem solving (Cooper *et al.*, 2008; Cooper and Sandi-Urena, 2009); an instrument that allows faculty to measure the extent to which laboratory activities include inquiry based learning (Fay *et al.*, 2007); an instrument that measures student attitudes about learning chemistry via semantic differentiation (Bauer, 2008); and content assessments that include student estimates of item complexity (Knaus *et al.*, 2009).

With this broad range of instrument development it is now possible to consider a new working paradigm, in which curriculum development and assessment development occur symbiotically. This model is depicted in Fig. 1.

Most meaningful reform in the teaching and learning of any field begins at the top of this diagram, with an attempt to change some specific aspect of the curriculum or teaching environment. Quite often, the result of such a change shows only limited effectiveness with traditional, well-established assessments such as standardized tests, because these tests typically address traditional course content and skills.

[a] *Department of Chemistry, Iowa State University, Ames, IA 50011, USA; e-mail: taholme@iastate.edu*
[b] *Department of Chemistry and Biochemistry, Miami University, Oxford, OH 45056, USA; e-mail: bretzsl@muohio.edu*
[c] *Chemistry Department, Clemson University, Clemson SC 29634, USA; e-mail: cmelani@clemson.edu*
[d] *Department of Chemistry, University of South Florida, Tampa, FL, 33620 USA; e-mail: jlewis@cas.usf.edu*
[e] *Center for Assessment, PO Box 351, Dover, NH, 03821, USA; e-mail: ppaek@nciea.org*
[f] *Department of Chemistry, University of Iowa, Iowa City, IA 52242, USA; e-mail: norbert-pienta@uiowa.edu*
[g] *Department of Chemistry, University of California-Berkeley, Berkeley, CA 94720, USA; e-mail: astacy@calmail.berkeley.edu*
[h] *IMMEX Project, UCLA / Learning Chameleon, 5601 W. Slauson Ave, #255, Culver City, CA 90230, USA; e-mail immexr@gmail.com*
[i] *Department of Chemistry, Purdue University, West Lafayette, IN 47907-2084, USA; e-mail: mtowns@purdue.edu*

This journal is © The Royal Society of Chemistry 2010

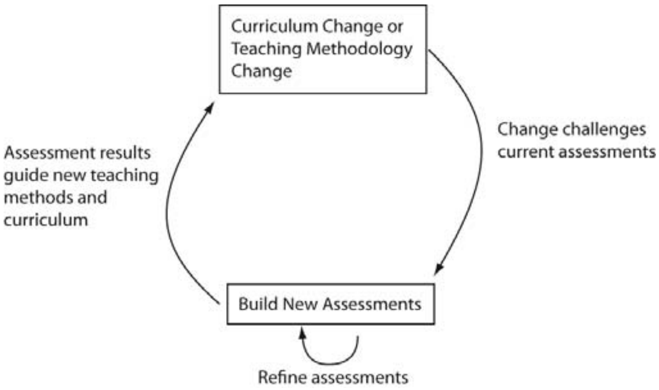
Curriculum Change or Teaching Methodology Change
Change challenges current assessments
Build New Assessments
Refine assessments
Assessment results guide new teaching methods and curriculum

Fig. 1 Schematic depiction of the curriculum/assessment development system for data-enhanced decision making in chemistry education reform.

Changing a course or a curriculum without a measureable improvement in or a curriculum without a measureable improvement in outcomes is not altogether appealing, so in many cases the instructor who has made the changes seeks out other ways to measure differences. Such individually created surveys may provide information that students 'like' the new activity or style of teaching. However, this subjective measure is not generally useful in devising ways to move forward with further improvements, nor is it useful in determining whether students have actually learned particular concepts and skills. A more meaningful alternative would be to identify characteristics of student learning likely to be affected by the change in pedagogical approach or curriculum, and specifically devise and validate new assessments to measure that aspect of learning. Thus, in this case, the teaching methodology or curricular change drives the development of assessment, and the results of the assessment help establish that meaningful reform has occurred.

Once developed, quality assessments can be applied to other environments both in other courses and with other pedagogies. Thus, the process by which quality assessment is disseminated plays a large role in continuing to drive forward the cycle of assessment-enhanced reform of courses or teaching. In many cases, the application of a new assessment instrument in a new institution, or a new course within a single institution, leads to unexpected discoveries. If student completion of general chemistry courses, for example, leads them to have expectations that diverge further from those of the chemistry faculty (Grove and Bretz, 2007) than before they enrolled in the course, such data should cause some re-evaluation of the activities or structure of the general chemistry coursework. This type of data-driven decision about coursework or teaching methods can then catalyze the adoption or adaption of new methods in the course, which reinitiates the cycle.

This iterative model provides a mode of operation for curriculum and teaching reform that may seem appealing, but barriers remain to its wide-scale implementation. In particular, quality assessments that address multiple factors related to learning may not be available or accessible to curriculum developers, instructors and researchers interested in reforming how and what chemistry is taught. This barrier is explicitly being addressed by a collaboration among a number of researchers at several schools, as will be described further here.

A premise: assessments beyond content knowledge can be useful

A defining feature of the model depicted in Fig. 1 lies in the fundamental premise that educational reform is legitimately informed by multiple modes of assessment.(Cooper *et al.*, 2010) In particular, while careful attention to content knowledge test development is vital to educational measurement, there are additional aspects to knowledge that merit attention as well. From the perspective of measurement theory, assessing content knowledge has a built-in advantage in most course frameworks because instructors routinely use multiple measures (*e.g.* mid-term exams and final exams). Errors in measurement, which are just as fundamental in educational measurement as they are in the chemistry laboratory, are less likely to have a profound effect on outcomes when the measures are repeated, particularly throughout the course. Indeed, the primary factor that limits the introduction of even more content measures (via graded homework or testing) is typically time rather than any sense that more measurements would not be valuable. However, many assessments measure content knowledge in the absence of any other skills (problem solving, ability to frame a scientific question, ability to transfer knowledge to novel situations, etc.); that is: factual recall is much easier to assess than the ability to analyze, evaluate, or synthesize data, skill development, or measures in the affective domain.

Assessment of these dimensions of student learning beyond content knowledge also takes time, and as a result, instruments that measure these constructs are likely to be used relatively infrequently. This means that the need to have well developed and validated instruments is even greater than for content assessments. If the measurement will be made only once in a given course, it is exceptionally important that the instrument used provides reliable data. Given this common constraint, the challenge of instrument development becomes clear. Significant time and expertise are required to devise a useful instrument and carry out the research required to measure its validity and reliability. Thus, it makes sense to identify in advance what qualifies as useful information for the purpose of devising educational improvements. While a number of ways exist to approach this task, probably the most fruitful method is to consider theories of learning[2], use them to identify factors that either enhance or inhibit student learning, and then devise ways to measure those factors accordingly.

For example Novak developed and refined Ausubel's theory of meaningful learning. (Ebenezer, 1992; Novak, 1998; Bretz, 2001) This theory delineates three dimensions of learning, namely cognitive (where content knowledge growth occurs), affective (where student attitudes change) and psychomotor (where physical skills or performance aspects are gained.) An assessment program that provides insight into all three dimensions of student learning will provide access to

Table 1 Assessment methods for various aspects of learning

Theory base		
Cognitive (content)	Affective	Psychomotor
ACS exams	Semantic Differential (Bauer, 2008)	IMMEX (also cognitive) (Stevens and Palacio-Cayetano, 2003)
ChemQuery (Claesgens *et al.*, 2008)	MSLQ (Pintrich and Johnson, 1990)	CHEMX (also cognitive) (Grove and Bretz, 2007)
Concept inventories (Hestenes *et al.*, 1992; Mulford and Robinson, 2002)		MCA-I (also cognitive) (Cooper *et al.*, 2008)
ROT (Bodner and Guay, 1997)		TOLT (also cognitive) (Tobin and Caple, 1981)
		GALT (also cognitive) (Roadrangka *et al.*, 1982)

a more robust and more nuanced view of student progress. Such enhanced assessment is far more likely to capture multiple aspects of student learning gains associated with curricular or teaching reform than measuring only content knowledge gain. Table 1 provides a classification of assessment methods into these categories.

A second premise: time is precious – use Occam's Razor in assessment instrument development

When instruments are first developed, it is rarely obvious immediately how long or extensive they should be. Consequently, there is a tendency to devise a large number of items in order to be confident that no aspect of the desired measure is underspecified. This process is predictable for most new instruments, but ultimately it is essential to pare the instrument back to use as few items as possible to reliably measure what is desired. In this sense, the Occam's Razor (Baker, 2007) test becomes vital for the ultimate success of an instrument – if success is defined in terms of usability in the 'real world' environment of classroom instruction, rather than in a research study. This collaboration has designed and carried out multiple examples of how this process can be envisioned in the past year.

For example, the semantic differential instrument devised by Bauer (2008) initially included twenty items that measured student attitudes (within the affective domain of the operative theory model.) The instrument was validated within several classes of students and found to provide useful information in this domain. Lewis and colleagues (Xu and Lewis, 2010) carried out further analysis, including exploratory factor analysis, and found that a substantial fraction of the variance in the data derived from the instrument could be explained by two factors, and only eight items were required to quantify these factors. Thus, a pared-down semantic differential instrument was devised and placed in the field. Results of this new instrument are methodologically akin to those measured with the longer instrument, yet the new instrument takes only a few minutes to administer, thus leading to more environments in which it might be utilized.

A second example currently being explored to reduce the time needed to administer an instrument is underway with the metacognitive awareness instrument (MCA-I) from Cooper and colleagues.(Cooper, Sandi-Urena *et al.*, 2008; Cooper and Sandi-Urena, 2009) The original instrument includes items that were phrased with both positive and negative connotations. Not surprisingly, these two categories arose as largely independent factors in this instrument. Thus, an experiment is now underway to determine if using only the positively stated items will provide the same level of reliability in measurement. If successful, this research will enable instructors to measure metacognitive awareness in less time. In addition other metacognitive instruments (Grove and Cooper, 2010) are being evaluated in the same way, to produce a much shorter instrument.

It is also possible to devise measures that meet the Occam's Razor test from the outset. While it is difficult to measure the validity and reliability of a single item, it may be possible with properly designed items to obtain such estimates of quality for as few as three items. Thus, triads of questions for use with student response systems in the classroom have been developed and administered with large groups of students on many campuses. For example, the concept that breaking a chemical bond is an endothermic process can be couched in different contexts or with varying quantitative precision, to form a triad of items, all of which have this fundamental chemical concept at their core. This type of development holds the promise of establishing validity for the items while simultaneously informing teaching. Validity in this context, however, remains challenging to establish, as item order effects (changes in what a specific clicker question measures dependent on when that question is tested relative to the others) are difficult to avoid, for example.

Devising assessments for cognitive skills that span content domains

One key aspect of content assessment is that it typically embeds the evaluation of skills within items that are designed to measure specific content. Thus, if there is a desire to measure critical thinking, or problem solving skills, the tendency is to use content-based exercises as a proxy for this style of measurement. While open-ended responses for exercises in chemistry are capable of providing insight into student approaches for a particular exercise, the challenge of generalizing this assessment remains an important one.

One approach that has established a capacity to measure these generalizable skills is the IMMEX system for measuring problem solving strategy growth. The typical protocol is for students to encounter a complex, open-ended problem in a web based environment. Students are free to choose from a sizable array of information that is available to help to devise a solution to a problem solved, and importantly, all actions are stored in a database. With enough student performances, data-mining methods, such as Artificial Neural Networks (ANNs) and Hidden Markov Models (HMM) can be applied to identify clusters of similar strategies (Stevens and Palacio-Cayetano, 2003). Thus, strategies used to solve the problems can be categorized and described. In addition, the difficulty of the specific version of a complex problem (a typical

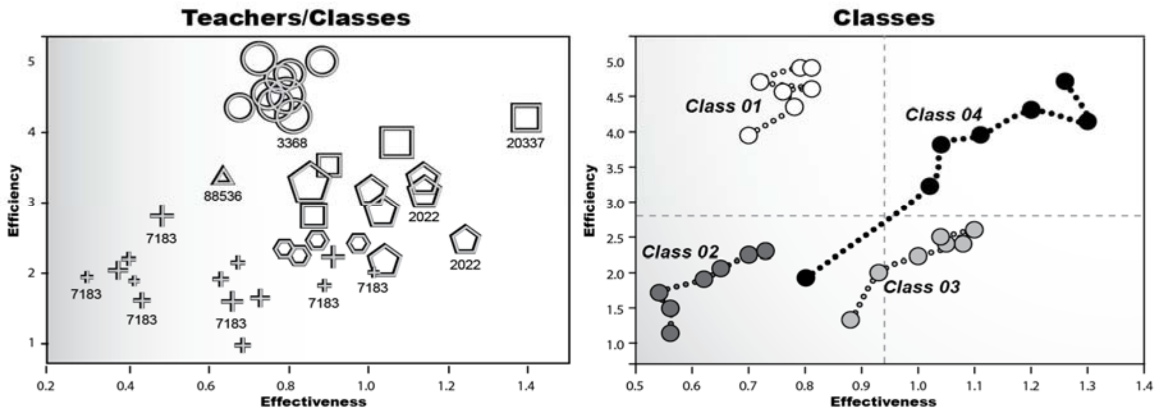

Fig. 2 Aggregated efficiency and effectiveness measures of classrooms that performed Hazmat. A) The dataset was aggregated by teachers (symbols and text) and class periods with the efficiency (scale 0-6) and effectiveness (scale 0-2) measures calculated as described elsewhere (Stevens and Thadani, 2007). The symbol sizes are proportional to the number of performances. B) The Efficiency/Effectiveness measures are stepwise plotted for seven Hazmat performances for four representative classes. The axis is bisected by dotted lines indicating the average efficiency (2.78) and effectiveness (0.96) measures of the dataset creating quadrant combinations of high and low efficiency and effectiveness.

IMMEX problem has between ten and thirty versions that are referred to as ‘clones’) can be established using Item Response Theory (IRT), which provides a second measureable dimension – namely effectiveness.

The combination of these two aspects of problem solving gives rise to a measure for aggregating efficiency and effectiveness of problem solving, as shown in Fig 2. This depiction provides quadrants, where performances that lie in the upper right quadrant, for example, are both efficient and effective, arguably the most desirable problem solving state for a student to achieve. This measure can be aggregated on any number of levels, including that for an individual student, for a single class, or for all classes of a particular instructor. This data format provides useful information for the instructors to make changes to their teaching, such as focusing on specific students who are making similar errors. Data can be obtained for any number of problem solving scenarios, and the data presented here is derived from a problem called Hazmat, which is essentially a general-chemistry level qualitative analysis scenario.

Looking at Figure 2, several levels of understanding may be derived by using this system. On the left side of this figure the efficiency and effectiveness of problem solving is plotted for multiple classrooms of different instructors. The different classes (shown by the same shape) often cluster in the same area of the efficiency / effectiveness continuum (for instance the squares to the upper right, or the +’s to the lower left) suggesting a consistent influence of the instructor on student’s problem solving outcomes. Such comparisons may serve as an assessment driven stimulus to consider teaching methodologies that prove to be successful. The figure to the right plots trajectories for four classes showing how problem solving improved as an increasing number of Hazmat cases were attempted. The progress towards efficient and effective problem solving made in ‘Class 04’ is dramatic, and may suggest that the particular pedagogies used by the teacher in this class are potentially worth disseminating.

Acquiring this type of formative assessment data and delivering it rapidly to instructors could indicate if and when the instructor needs to intervene in student practice. Importantly, this construct is transferable across problems, which opens the possibility of tracking students throughout a course to see how their problem solving skills evolve during the span of a semester. Work based on this concept involves using IMMEX as a form of automatically graded homework, and the assignment of several problems at some interval. This strategy for helping students become better problem solvers (and measuring the success of the strategy) is ongoing at this time with students enrolled in a chemistry for engineering students course. Students carry out a different IMMEX problem every three weeks, and solve-rates for the problems generally improve, even though the later problems are built from more demanding chemistry content (Caruthers and Holme, 2010).

Another key strategy for content domain-spanning assessment is the development of concept inventories. Ideally, items in such inventories are devised to elicit student understanding of key, typically broad-based, concepts. The success of such inventories in physics (Hestenes *et al.*, 1992) has led to a number of attempts to devise similar instruments in other domains, including chemistry (Mulford and Robinson, 2002). Work remains to establish how best to use this form of assessment so that it does not become an exercise where students learn the ‘right’ answer rather than the underlying concept, for example.

Finally, it is important to realize that learning theories suggest that there are overarching cognitive progressions through which students must pass. It is possible to devise assessments with this cognitive dimension in mind. The ChemQuery program does this explicitly (Claesgens *et al.*, 2008), and by adjusting the content knowledge assessments to include this cognitive development, it provides detailed information about the success of the curriculum being followed. Specifically, a model for the cognitive hierarchy related to content knowledge has been devised beginning with ‘notions’ that need not even be couched within a scientific vocabulary. The next level is ‘recognition’, where students begin to use scientific language followed by ‘formulation’ where students begin connecting more than one scientific concept. A level designated 'construction' finds students using fully developed scientific models, and the final level 'generation' requires that students be able to identify and research questions that would extend such models. Progress towards the higher levels of this hierarchy is the goal of most chemistry courses, and assessments can be constructed to measure that progress.

Assessment-based educational choices

The final piece of the puzzle for data-enhanced educational reform efforts lies in the ability to utilize data to make wise choices for new teaching methods or curriculum changes. There are clearly examples in each of these categories.

First, the use of multiple assessments, along with targeted teaching interventions is capable of establishing the relative effectiveness of teaching methods. For example, Cooper and colleagues (Cooper, Cox *et al*., 2008) have established, using IMMEX, that students will stabilize on a problem solving strategy relatively quickly, and hold to that strategy with some tenacity. In other words, students will retain their strategy even if it does not have a high rate of success. They further established that working in collaborative groups provides a useful method for moving students towards more productive strategies in subsequent individual problem solving attempt. Other studies by Cooper and Sandi-Urena (2009) showed that the IMMEX strategy, IRT ability and MCA-I score were correlated, and that for many students these measures could be used as a proxy for each other.

This triad of assessments can now be used to show improved outcomes arising from interventions based on research on teaching and learning, and these outcomes may be difficult to evaluate using traditional methods. For example, Sandi-Urena and Cooper have developed an intervention designed to help students become more metacognitive problem solvers (Sandi-Urena and Cooper, 2010). It was found that students who participated in this activity had increased levels of metacognition (from their change in MCA-I score) and better problem solving skills, as measured by IMMEX measures (ability and strategies on unrelated IMMEX problems), than students in the control group, who did all of the activities of the experimental group *except* the metacognitive intervention. Another study looked at the effect of a cooperative lab-based program (Cooper, 2009) on student problem solving (Cooper *et al.*, 2010). The laboratory environment requires students to plan, monitor, and evaluate their activities as they design experiments to solve problems, and analyze their data. Outcomes from this kind of laboratory program are difficult to assess using measures of student learning such as course exams, since any gains would be most likely to occur in problem solving and decision making. Students who participated in this laboratory program were found to have significantly better problem solving abilities, and metacognitive levels than the control group. These two studies have given evidence that would otherwise be difficult to obtain for *research based* educational methods, and have in fact shown that these are now *research validated* educational methods.

Second, assessments used by many groups of students with different backgrounds and/or educational goals, can identify challenges for content learning that are robust, regardless of the specifics of student populations. An example of this is derived from work done by Pienta and co-workers (Pienta, 2010) who have studied the role of cognitive load (Sweller, 1988). in student problem solving success. In this work, it is found that common operations in chemistry, such as unit conversions, are actually harder for students to carry out successfully in one direction than the other. In other words, student errors arise more frequently in converting from mL to L than in converting L to mL, and this type of error distribution arises in any level of student constituency, from Preparatory Chemistry (for students who are not prepared for standard college chemistry) to Chemistry for Engineering Students, who typically have relatively strong math backgrounds. That this observation is made for seemingly low difficulty math skills suggests that changing the manner in which these skills are taught might be worth considering when devising changes in the curriculum for entry-level college courses.

Finally, it is possible to establish content-based assessments that show the impact of the introduction of different content in a course. Knaus, *et al.* (2010) established with a combination content/affective measure that the use of examples of nano-science within a course lead to gains in the efficiency of student learning for content within this broad category. These gains are not dependent on direct instruction on the material, but rather show that any introduction to the new content area, in this case nanotechnology, leads to student gains in learning efficiency, within that field. For example, students in a course that included roughly 50 minutes of total instruction time in nanoscience (out of roughly 2000 minutes of instructional time), performed significantly better and with greater efficiency on unfamiliar nanoscience items than students in a course that had no instructional time devoted to the topic at all. Thus, by using an assessment that is sensitive to more than just the content knowledge, data about the efficacy of curricular choices (the inclusion of nanoscience in a general chemistry course) can be established.

Summary

The premise of this project is that assessment carried out with multiple measures provides a model that can better inform reform efforts in chemistry education. This premise is based on the experience of the authors in collaboration over many institutions, using a large number of assessment instruments. This model calls for a new synergy between curricular development and assessment development. The ability to measure aspects of learning beyond specific content knowledge that may be tested in a traditional manner is particularly important within this model. For example, it could be argued that the art of problem solving in education is underdeveloped despite a large research effort to understand problem solving. The difficulties are attributable, in part, to the primitive state of problem solving assessment. It is much easier to test for the facts of science than it is to test for the other critical types of science understanding which has ramifications for how science is taught.

There is a critical need, and increasing calls for the development of new assessments for scientific reasoning and problem solving (Jonassen, 2007; Alberts, 2009). Thus, a model for education reform that includes assessment development as a fundamental component shows promise for tackling particularly challenging aspects of improved teaching and thus student learning.

Acknowledgements

We acknowledge funding for the collaborative project derived from the National Science Foundation, (DUE-0817257, 0817279, 0817297, 0817409, 08117594.)

Notes and References

[1] For information about the Exams Institute see: http://chemexams.chem.iastate.edu/

[2] Several articles address theories in on-line only form at *J. Chem. Educ.*, 2001, **78**, 1107.

Alberts B., (2009), Redefining science education, *Science*, **323**, 437

A. Baker, (2007), "Occam's Razor in science" a case study from biogeography, *Biol. Philos.*, **22**, 193-215.

Barbera J., Adams W. K., Wieman C. E. and Perkins K. K., (2008), Modifying and validating the Colorado Learning Attitudes about Science Survey for use in chemistry," *J. Chem. Educ.*, **85**, 1435-1439.

Bauer C. F., (2008), Attitude towards chemistry: a semantic differential instrument for assessing curriculum impact, *J. Chem. Educ.*, **85**, 1440-1445.

Bodner G. M. and Guay R. B., (1997), The Purdue visualization of rotations test, *Chem. Educator*, **2**, 1-18.

Bretz S. L, (2001), Novak's theory of education: human constructivism and meaningful learning, *J. Chem. Educ.*, **78**, 1107; DOI: 10.1021/ed078p1107.6.

Caruthers H. and Holme T., 2010, unpublished work.

Claesgens J., Scalise K., Wilson M. and. Stacy A. M, (2008), Assessing student understanding in and between courses in chemistry, *Assessment Update*, **20**, 6-8.

Cooper M.M., (2007) Data-driven education research, *Science*, **317**, 1171.

Cooper M. M., (2009), Cooperative chemistry laboratories, 4th Ed., McGraw-Hill, New York, NY.

Cooper M. M., Cox Jr. C. T., Nammouz M., Case E. and Stevens R. H., (2008), An assessment of the effect of collaborative groups on students' problem solving strategies and abilities, *J. Chem. Educ.*, **85**, 866-872.

Cooper M. M. and Sandi-Urena S., (2009), Design and validation of an instrument to assess metacognitive skillfulness in chemistry problem solving, *J. Chem. Educ.* **86**, 240-245.

Cooper M. M., Sandi-Urena S., Gatlin T., Bhattacharyya G. and Stevens R., (2010), Mixed methods study: effect of cooperative problem based lab instruction on regulatory metacognition and problem solving skills and performance, *Sci. Educ.*, submitted.

Cooper M. M., Sandi-Urena S. and Stevens R., (2008), Reliable multi method assessment of metacognition use in chemistry problem solving, *Chem. Educ. Res. Pract.*, **9**, 18-24.

Ebenezer J. V., (1992), Making chemistry more meaningful, *J. Chem. Educ.*, **69**, 464-467.

Fay M. E., Grove N. P., Towns M. H. and. Bretz S. L, (2007), A rubric to characterize inquiry in the undergraduate chemistry laboratory, *Chem. Educ. Res. Pract.*, **8**, 212-219.

Grove N. P. and Bretz S. L., (2007), CHMEX: Assessing students' cognitive expectations for learning chemistry, *J. Chem. Educ.*, **84**, 1524-1529.

Grove N. and Cooper M. M., (2010), unpublished work.

Hestenes D., Wells M. and Swackhamer G., (1992), Force concept inventory, *Phys. Teach.*, **30**, 141-158.

Jonassen D. H., ed., (2007), *Learning to solve complex scientific problems*, Lawrence Earlbaum Associates, New York, NY.

Knaus K. K., Murphy K. L. and Holme T A., (2009), Designing chemistry practice exams for enhanced benefits, *J. Chem. Educ.*, **86**, 827-832.

Knaus K. K., Murphy K. L. and Holme T. A., (2010), The impact of nanoscience context on multiple choice chemistry items, in *Nanotechnology in Undergraduate Education*, K. A. O. Pacheco, ed., ACS Symposium Series, Oxford University Press, pp. 7-18.

Labov J. B., (2007), The intersection of STEM assessment, accountability, and education policy: a 'gathering storm' for higher education?, in *Proceedings of the National STEM Assessment Conference*, D. Deeds and B. Callen, eds., National Science Foundation, Washington, DC,. pp. 3-11.

Mathew J. M., Grove N. P. and Bretz S. L., (2008), Online Data collection and database development for survey research in chemistry education, *Chem. Educator*, **13**, 190-194.

Mulford D. R. and Robinson W. R., (2002), An inventory for alternate conceptions among first-semester general chemistry students, *J. Chem. Educ.*, **79**, 739-744.

Novak J. D., (1998), *Learning, creating, and using knowledge*, Mahwah: Lawrence Erlbaum,.

Pienta N. J., (2010), unpublished work.

Pintrich P. R. and Johnson G. R., (1990), Assessing and improving students' learning strategies, *New Directions for Teaching and Learning*, **42**, 83-92.

Roadrangka V., Yeany R. H. and Padilla M. J., (1982), *GALT. Group test of logical thinking*, University of Georgia, Athens, GA.

Sandi-Urena S. and Cooper M. M., (2010), Enhancement of metacognition use and awareness by means of a collaborative intervention, *Int. J. Sci. Educ.*, in press.

Soller A. and Stevens R. H., (2007), Applications of stochastic analysis for collaborative learning and cognitive assessment, in *Advances in latent variable mixture models*, G. Hancock and K. Samuelson, eds., Information Age Publishing, pp. 217-254.

Stevens R. H. and Palacio-Cayetano J., (2003), Design and performance frameworks for constructing problem-solving simulations, *Cell Biol. Educ.*, **2**, 162-179.

Stevens R. H., Soller A., Cooper M. and Sprang M., (2004), Modeling the development of problem solving skills in chemistry with a web-based tutor, in *7th International Conference Proceedings*, J. C. Lester, R. M. Vicari, and F. Paraguaca, eds., Springer-Verlag Berlin, Heidelberg, Germany, pp. 580-591.

Stevens R. H. and Thadani V., (2007), Quantifying students' scientific problem solving efficiency, *Instruction, Cognition and Learning (TICL*), **5**, 325-337.

Sweller J., (1988), Cognitive load during problem solving: effects on learning, *Cogn. Sci.*, **12**, 257-285.

Tobin K. G. and Caple W., (1981), The development and validation of a group test of logical thinking, *Educ. Psychol. Meas.*, **41**, 413-23.

Towns M. H., (2010), Developing learning objectives and assessment plans at a variety of institutions: examples and case studies, *J. Chem. Educ.*, **87**, 91-96.

Xu X. and Lewis J., (2010), Refinement of a chemistry attitude measure for college students, *J. Chem. Educ.*, submitted.