Query Details

Back to Main Page

There is no Author Query !

Aligning Assessment Goals with the Current and Future Technologies Needed to Achieve Them

Melanie M. Cooper
Email : mmc@msu.edu
Affiliationids : Aff1, Presentaffiliationid : Aff1

Michael W. Klymkowsky ✉
Email : michael.klymkowsky@colorado.edu
Email : klym@colorado.edu
Affiliationids : Aff2, Correspondingaffiliationid : Aff2

Aff1 Department of Chemistry, Michigan State University, East Lansing, MI, USA

Aff2 Department of Molecular, Cellular Developmental Biology, University of Colorado Boulder, Boulder, CO, USA

Abstract

The issue of assessment involves two interdependent drivers: the purpose(s) of the assessment and how such assessments can be applied at scale, that is, in large and, in some cases, remote settings. The simplest assessment goal, to sort students by what content they know or can recognize as correct, often involves a variety of "forced-choice" or fill in the blank questions that are readily analyzed by computers. Higher-level assessments that evaluate the extent to which students can access and apply their knowledge to new situations (as opposed to remembering previously presented examples), and can be used to develop students' working knowledge, demand more sophisticated Socratic approaches aimed at making student presumptions explicit, together with their relevance and implications. Progress along these lines involves the automated analysis and response to drawn responses (graphs and such), as in the beSocratic™ system. Future extensions will require an iterative feedback system that can analyze students' textual responses "on the fly" and pose disciplinarily relevant and clarifying Socratic questions. We consider the current state of affairs in achieving this goal.

Keywords

Assessment design
Formative assessment
Constructivism
Instructional technology
Socratic prompts

8.1. Introduction

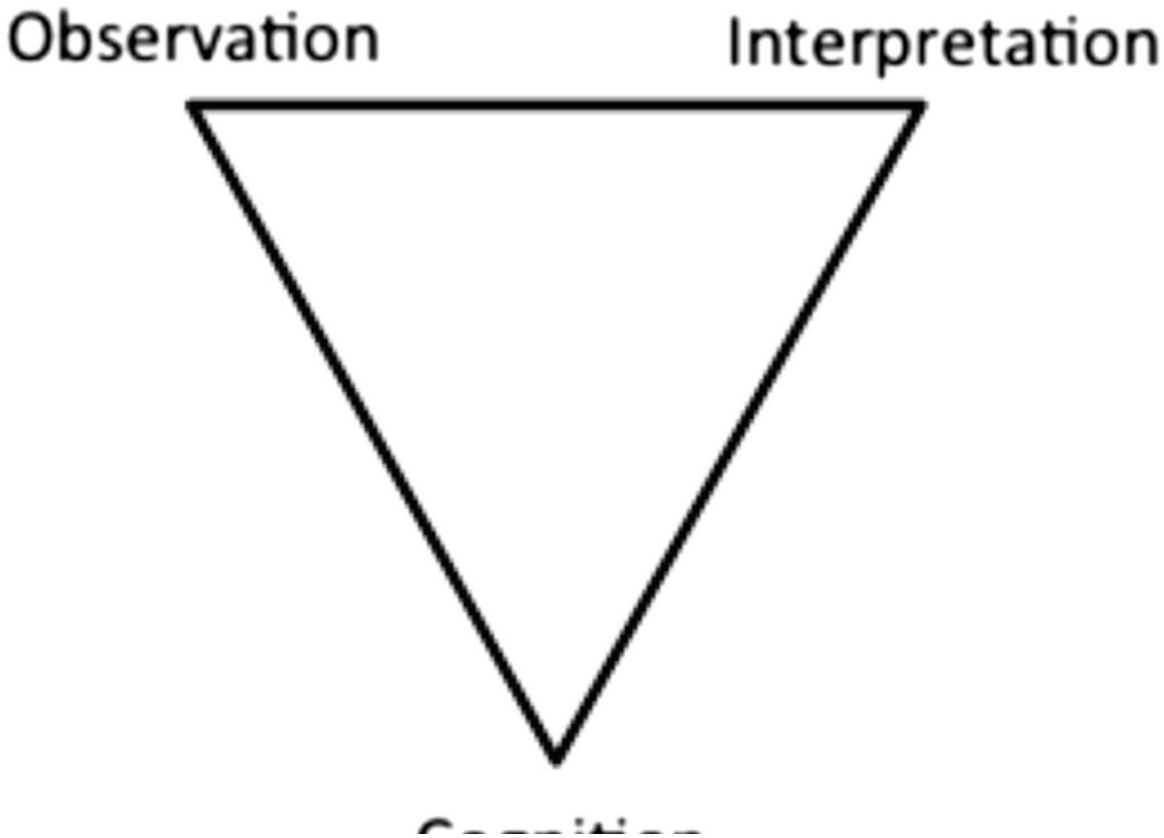
Knowing what students know is a difficult proposition (National Research Council, 2001). While it is relatively easy to ascertain whether students are able to remember facts, apply rules or heuristics, or solve algorithmic problems,

understanding how students' reason about more complex ideas is much more difficult. Assessment of simple recall and algorithmic problems can readily be achieved by multiple-choice tests that are easily administered and machine graded. Such approaches have become commonplace in many areas of higher education, particularly where large numbers of students are involved. However, we know that the ways that students are assessed may impact the choices that students, and instructors, make about teaching and learning. We can assume that students are rational actors; if they see that their grades will be determined solely by whether they can remember facts and perform rote calculations, then that is what they will learn to do. Indeed, the adage "If you don't assess what is important, then what's assessed becomes important" captures this reality (Miller & Parlett, 1974). Such recall and algorithmic (heuristic)-based assessments can lead to fragmentation and trivialization not only of what students know but of how curricula are designed and delivered so that outcomes can be assessed in this way. That being said, it is unlikely that most faculty would place memorization and the rote application of mathematical formulae as their goals, which leads us to the central problem—how to assess student reasoning on more complex tasks, at scale, without increasing costs in terms of time and personnel. In this chapter we discuss ways in which technology can be harnessed to address this problem.

Designing Assessments That Elicit Evidence of Learning We start by considering theoretical aspects of assessment design. Perhaps the most influential approach was described in the National Academy of Sciences publication *Knowing What Students Know* (National Research Council, 2001). The authors of this consensus report proposed that assessments be considered as evidence-driven arguments. That is, assessments should elicit evidence that can be used to support an argument about what students know and are able to do. A visual representation of this process is known as the assessment triangle (Fig. 8.1), where the three vertices represent (1) the model of cognition that supports assessment design; (2) the observations that the assessment design is expected to elicit; and (3) how those observations will be interpreted. For example, if we want to know what students can memorize or recognize, we would design very different test items than if we are interested in whether students are able to use their knowledge and apply it to new situations.

Fig. 8.1

The assessment triangle



Cognition

8.1.2 Models of Cognition Such differences stem from the model of cognition that guides assessment development. Over the years, many models of learning have been proposed (National Academies of Sciences, Engineering, and Medicine, 2018; National Research Council, 1999) and with them concomitant approaches to designing learning environments and assessments. It is beyond our scope here to review them all, but in general and over time, there has been a paradigm shift away from behaviorist learning models (Gagne, 1965; Skinner, 1954) such as direct instruction and associated measures of learning based on what we will call traditional tests, tests often based on right/wrong and forced-choice answers, factual recall, and the application of procedures and equations. Over time, most researchers have moved to more constructivist perspectives on learning and assessment. That is, instead of viewing knowledge as something that is transferred from instructor to student, learning is viewed as an act of construction by the student, through a process aided by the instructor and shaped by the design of the learning materials. Interestingly, as pointed out by Shepard, the shift from a behaviorist to a constructivist learning model is often *not* accompanied by a shift in assessment design and use (Shepard, 2000). Too often constructivist pedagogical approaches, based on active engagement of students with specific terms and concepts, and their application, are not accompanied by constructivist rethinking of course design or assessments. For example, the highly cited report on the efficacy of "active learning" strategies, as opposed to "passive" lectures (Freeman et al., 2014), relies almost exclusively on multiple-choice assessment instruments known as concept tests or concept inventories that, as we will discuss, are often incompatible with constructivist learning theories.

Even within constructivist learning models, there are an array of variants as to just how knowledge is constructed, that is, how learning occurs, and how students' prior knowledge is integrated (and modified) in the course of their development of a coherent understanding of disciplinary ideas and their application. At the same time, there has been a shift from models that focus on replacing students' incorrect ideas (misconceptions) with correct ideas (Posner et al., 1982) to viewing student ideas as resources that may or may not be useful, or even be recognized as relevant in a particular situation. Through iterative assessment and associated feedback, commonly termed formative assessment, students' ideas can be reshaped and woven together to construct a disciplinarily coherent and applicable framework (Hammer, 2000). Different models of, and approaches to, learning require (or are satisfied ~~with~~by) different assessment strategies.

8.1.3 Assessment: From Simple to Complex Here we consider how various assessments align with different cognitive models (or not) and the evidence of learning that they have the potential to elicit. If we are simply interested in whether a student has memorized a fact or is able to use an equation to calculate a numerical answer, assessment design can be quite simple. Typically, such questions are written so that there is a single unambiguously correct response. Consider the question "in what molecule is genetic information stored in organisms?" The answer is DNA. A multiple-choice or fill in the blank response will suffice to determine whether a student "knows" or can recognize the correct answer. As we discuss later, answering such questions does not provide us with any evidence that the student has a coherent knowledge framework that incorporates an understanding of where genetic information comes from, how it is stored, how it is used, or what features of DNA, and the cellular systems involved in its replication, repair, and access make it a uniquely suitable molecular system for information storage. We should not expect that students who are assessed through memorization/recognition-type questions will have developed the disciplinary expertise required. Such assessments are, however, easy to write and grade and can be administered online. Indeed, there has been a proliferation of publisher developed, textbook-associated test banks (as witnessed by a simple Google search). By their very nature, such test bank questions address surface-level knowledge. While fears of cheating (see Chap. 10 in this volume) can drive various anti-cheating strategies (such as randomizing question order), they do not address the underlying nature of the assessment instrument, what we might term its "resolution"—the level of understanding its use can reveal. Such exams rely on simplistic models about how people learn and can be counterproductive: supporting an overconfidence (a type of Dunning-Kruger effect) among both students and instructors that students' knowledge and instructional efficacy is much more robust than it actually is.[1] It is unlikely that students who are routinely tested in this manner will develop disciplinary expertise, that is, the ability to apply their knowledge to new and/or more ambiguous situations. For example, students who scored in the 75th percentile of the national average in the USA (using a multiple-choice test) have been shown to have considerable misunderstandings about important

ideas central to chemistry and biology (Williams et al., 2015). In fact, there is evidence that such forced-choice instruments tend to overestimate what students know (Lee et al., 2011) and that they tell us very little about what students can do with their knowledge. It should also be noted that such assessment items do not align with modern theories of cognition.

There are a few types of assessment instruments that are based on a model of cognition, have a research-based development protocol, and can be administered online and machine graded. These are what are now known as concept tests or inventories. Pioneered by Treagust (1988) and Hestenes et al. (1992), these multiple-choice tests are designed not to determine what students know, but rather to identify what "misconceptions" students may hold, that is, what ideas and presumption students incorrectly apply and that need to be addressed through instruction. There are now hundreds of such inventories, across many fields of study (Klymkowsky & Garvin-Doxas, 2020; National Research Council, 2012).

The model of cognition underlying concept tests and inventories, while acknowledging that students construct knowledge, implies that if we can identify misconceptions, they can be replaced or overwritten by instruction. Concept tests have had a major impact in physics, where students' understanding of the macroscopic world conflicts with established physical laws. Understanding why we are not projected into space by the Earth's rotation requires principles not readily deduced from everyday experiences.

In early work Strike and Posner (Posner et al., 1982) proposed that once a misconception is identified, instruction can induce cognitive dissonance that will lead students to become dissatisfied with their understanding—allowing the correct idea to replace the incorrect one. While this approach has been modified over the years, there is, in fact, little evidence that this is how learning actually works for the typical student or that this "replacement" approach brings about conceptual change, particularly for complex phenomena—which is the case for scientific processes, whether chemical, physical, or biological. Nevertheless, concept inventories are often used to monitor "learning outcomes" and to supply evidence for improved outcomes associated with specific educational strategies. In such cases, pre- and post-instruction test scores are used to calculate "learning gains" (Freeman et al., 2014; Hake, 1998), but what exactly has been learned is rarely, if ever, clearly stated, making such conclusions rather unconvincing.

Part of the problem lies with the multiple-choice format of concept test/inventories. While a student may choose the correct answer, we have no evidence about why that answer was chosen. Conversely, choosing an incorrect answer may not mean that the student has a misconception, but rather they may have misunderstood the question or are using reasoning and presumptions that the item writer did not intend. Certainly, if the distractors (wrong answers) are based on research on how students might answer that question, a student's choice of distractor can provide some insights into student thinking (Garvin-Doxas and Klymkowsky, 2008), but without further evidence, it is difficult to determine the cause of such problems. When distractors do not reflect actual student thinking (as is often the case when instructors construct the test items), students may more easily rule out implausible answers which can improve scores, while understanding may remain unaltered (Gierl et al., 2017). Another possibility is that such "lures" may actually implant ideas that students did not have previously (Butler et al., 2007; Butler & Roediger, 2008).

One final type of multiple-choice assessment that relies on extensive research on student thinking is the "ordered multiple choice" (OMC) instrument, which treats student learning as a progression of increasingly sophisticated ideas. These instruments were first introduced by Briggs and co-workers, as a way to provide effective and rapid diagnostic information about student learning. Because they are not a typical dichotomous (right/wrong) test, they provide more evidence about how students are thinking about a particular phenomenon (Briggs et al., 2006). OMC items are often based on learning progressions, which are "successively more complex ways of thinking about an idea that might reasonably follow one another in a student's learning" (Smith et al., 2006). In OMC instruments the distractors are developed from research on student thinking and are ordered in ~~increasing~~ levels of increasing sophistication using item response theory (IRT) (Embretson & Reise, 2013). That is, the answers are not right or wrong, but rather they differ in sophistication. Such OMC assessment instruments require a great deal of prior research and psychometric analysis before they are considered reliable and valid.

8.2.1 Constructivist Assessments for Constructivist Pedagogies and Learning Models Because we focus here on how technology can support the assessment of learning, we have so far considered the most common types of questions that lend themselves to automated delivery and grading on various instructional platforms. However, we hope that we have

convinced the reader that multiple-choice, fill in the blank, and true/false questions generally fail to provide convincing evidence about what students understand and perhaps even more importantly what they can do with their knowledge. Furthermore, these types of questions typically test only fragmentary knowledge, which does not align with constructivist models of how students learn. While we do not expect our students to become experts after 1 or 2 (or even 4) years of learning in a subject, our goal as instructors is to support students toward the development of knowledge that is useful and not inert—that is, toward a more expert-like understanding.

It is clear that experts' knowledge is structured differently from that of novices (National Research Council, 1999). Experts' knowledge cannot be reduced to sets of isolated facts; it is connected, conditionalized depending on circumstances, and reflects a deep understanding of the subject matter—the ability to recognize what is likely to be relevant and irrelevant to a specific situation. To develop such expertise requires that students are encouraged to construct more expert-like knowledge frameworks, a process that often involves multiple attempts, informed feedback (coaching), and repeated revisions. If we want to evaluate whether students are moving toward expertise, we must learn how to distinguish and evaluate more expert-like thinking. To do that, we must provide students with learning materials and formative and summative assessments that help them develop more connected, contextualized, and useful knowledge frameworks.

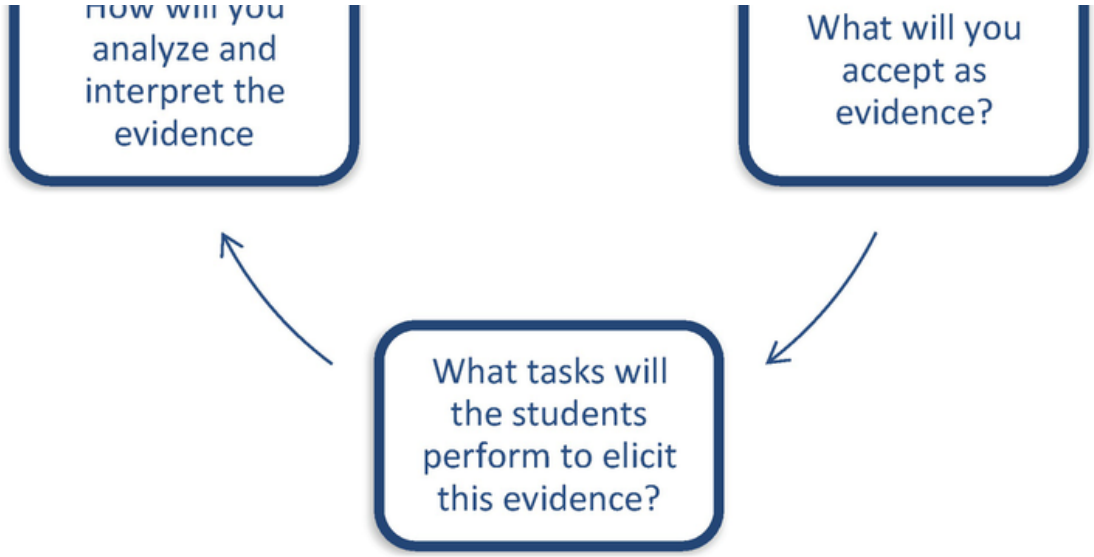
In much of our current work, we have adopted a constructivist model of cognition and used a resources framework (Hammer, 2000) through which we attempt to help students develop, recall, and use appropriate resources (which may be facts, skills, concepts, over-arching principles, and practices) that can be connected to produce a coherent and useful response to a prompt. In general, this requires that students are able to construct (rather than choose from a menu) their own responses, which consist of written explanations and arguments that may include or rely upon drawn diagrams, models, graphs, and pictures. In general, both modalities (written and drawn) are required to elicit evidence about what students know and can do. This, of course, poses a technological challenge; it is much more difficult to recognize and respond to students' constructed responses than it is to recognize the correct answer in a forced-choice instrument. However, if it is evidence that we are after, we must allow students to speak with their own words and drawings.

In our work we have used a modified version of evidence-centered design (ECD) (Mislevy & Riconscente, 2011) to design assessments that can elicit appropriate evidence. As shown in Fig. 8.2, ECD involves a number of steps:

1. Develop statements about what students should know and be able to do. This means explicitly acknowledging the model of cognition to be addressed.
2. Describe what evidence would convince us that students are able to use that knowledge in ways that we have specified.
3. Design tasks that have the potential to elicit such evidence.
4. Consider how the evidence elicited by the assessment will be analyzed and interpreted.

Fig. 8.2

Evidence-centered design



This is an iterative process, and, in practice, we find that it is typically necessary to repeat the cycle a number of times. Reframing of questions and prompts is also often needed to elicit informative student responses.

8.2.2 beSocratic: A Constructivist Teaching and Assessment System With these ideas in mind, let us consider the ways in which technology can support the delivery and evaluation of well-designed, beyond forced-choice, assessment tasks. Here we describe our work on the development and use of the beSocratic system in the context of courses in general chemistry, organic chemistry, and molecular biology. These are assessment tasks that have been administered to tens of thousands of students and so reflect a particularly promising approach.

beSocratic is a web-based assessment platform that allows students to complete tasks by drawing diagrams, graphs, and reaction mechanisms, by writing explanations and arguments, with minimal constraints (Bryfczynski, 2012; Cooper et al., 2014). That is, the evidence elicited by the task prompt is not constrained by the way the student must respond as is the case with forced-choice assessment systems. The system is designed so that many (although not all) types of diagrams and graphs can be automatically analyzed and responded to, but perhaps more importantly instructors can target known issues with student drawings and graphs and provide contextualized feedback that is increasingly specific, as necessary. An obvious extension of such a system would be the automated analysis and input-driven feedback to written responses. While not yet implemented, there are a number of promising strategies applicable to this task which are briefly discussed at the end of this chapter.

8.2.3 Examples of beSocratic Activities Here we describe a few beSocratic activities and the evidence they have provided about student learning. Non-covalent interactions or intermolecular forces (IMFs), that is, the forces that occur between molecules and between regions of larger molecules, are a fundamental idea in chemistry with significant implications for biological systems, relevant to molecular structure and properties and to how and why molecules interact (stick together) and come apart. Whether it be to explain phase changes, chemical reactions, enzyme-substrate binding, the effects of solvent and pH on the behavior of biomolecules and pharmaceuticals, or why geckos can walk up walls, the general principles are the same. These non-covalent interactions that go by a variety of names that include van der Waals interactions, London dispersion forces, and hydrogen bonding arise from the same phenomenon—nonuniform electron density distribution (whether permanent or induced) resulting in dipoles that lead to electrostatic attractions between molecules or between parts of large molecules.

In the first study, our goal was to determine whether students understood the nature of IMFs. The evidence we sought to elicit was simply "do students know what intermolecular forces are." We used beSocratic to record students responses to questions that asked them to describe their understanding of various types of IMFs, to draw three molecules (of ethanol), and to show where IMFs were operating (Cooper et al., 2015). Perhaps not surprisingly, students were able to provide the kinds of descriptions of IMFs that are prevalent in conventional textbooks, but to our consternation, these descriptions were

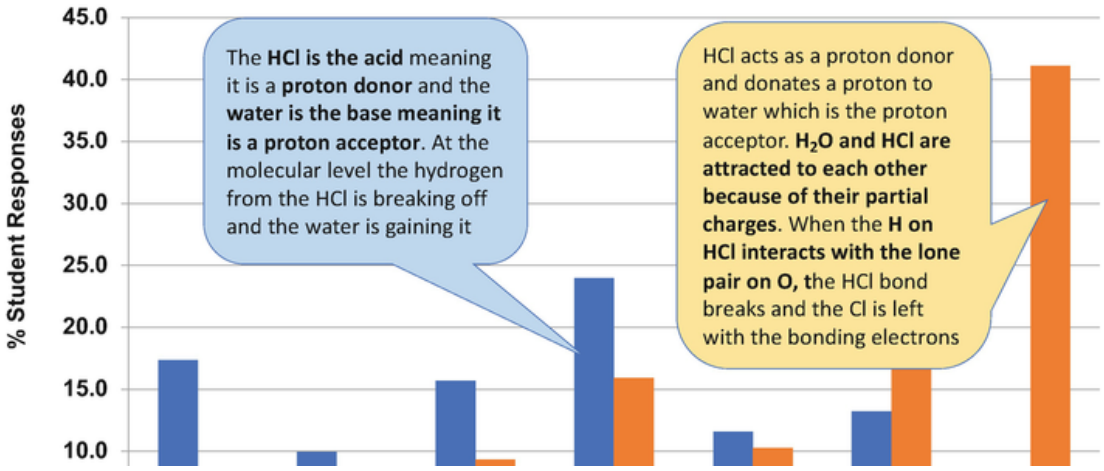
accompanied by drawings that in general tended to represent IMFs as forces operating within small molecules. This was true for all types of IMFs, and we have since replicated this study in many settings, with similar findings. Interestingly previous work, using multiple-choice instruments, failed to reveal such problematic responses (Schmidt et al., 2009). We also note that subsequent findings, in which students were enrolled in a transformed curriculum where the use of beSocratic was routine, showed marked improvements in student understanding of IMFs (Williams et al., 2015).

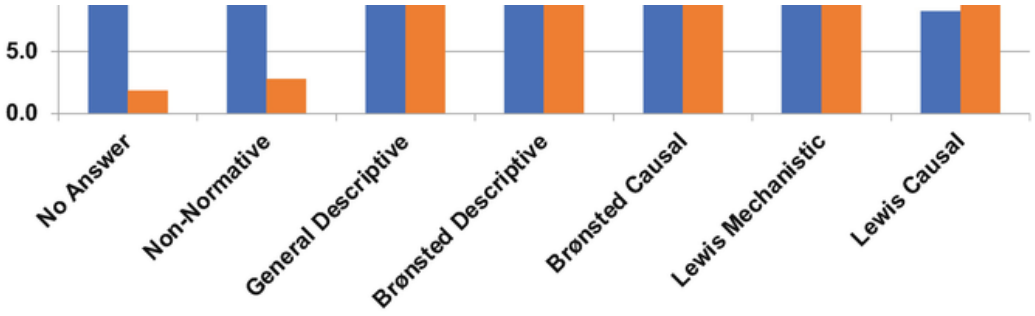
While such studies were revealing, what they did not elicit was any evidence about what factors (resources) students used to describe IMFs. Because IMFs are so central to understanding chemical and biological processes, we pushed further with activities designed to elicit more and better evidence about what resources students use and how they weave them together to provide explanations. Students were asked to write and draw explanations for what happens when two neutral atoms or molecules move toward each other; responses were captured in beSocratic and were coded by hand (Becker et al., 2016; Noyes & Cooper, 2019). The codes applied captured the type of reasoning that students used to answer the question and could be applied to both the written and drawn explanations. Student responses generally fell into three classifications of increasing sophistication. The lowest level corresponded to students who said the molecules are attracted but did not explain why. At level 2 students were able to leverage ideas about coulombic attractions—they were able to explain that the molecules were attracted because of opposite charges but did not provide a mechanism for how neutral molecules could have unequal charge distributions—which only occurred at level 3. We have replicated this study many times in many different institutions. Recently, working with the Automated Analysis of Constructed Responses (AACR) group (Urban-Lurain et al., 2015), we have been able to automate the coding of student responses (Noyes et al., 2020). This process is time-consuming and personnel intensive. Humans must code many responses to train the AACR system, but once trained the system is able to code student responses with the same accuracy as human coders. We return to the automated analysis of student responses shortly.

We describe one more example of a constructivist task designed to elicit the resources that students use as they reason about chemical phenomena to illustrate the importance of the prompt. In this study, we were interested in how students reason about acid-base reactions (Cooper et al., 2016). The methodology was similar to that described earlier: that is, student responses were characterized by the type of reasoning and the resources they used to address the prompt. Initially, we found that students' responses were at a rather low level. When students were asked to explain what was happening in a particular acid-base reaction, they provided descriptive responses about what acid-base reactions are and how acid-base reactions occur. By changing the prompt (Fig. 8.3) to make our expectations more explicit, much more sophisticated responses could be elicited from a similar cohort of students. Indeed, it has been our experience that it takes several attempts to design task prompts that elicit informative student responses and so provide clearer evidence of student thinking (Cooper et al., 2016). That is, we have had to go through the ECD assessment design cycle a number of times before we alight on a prompt that "works" to elicit a revealing response from the student.

Fig. 8.3

The change in student responses by changing the scaffolding in the prompt





One, perhaps not obvious, lesson has been that asking students to "explain" typically elicits disappointing results because students are not given sufficient guidance about what the instructor is looking for in the response, what exactly are they expected to explain. On the other hand, if the prompt is too scaffolded, the student is led to the appropriate answer, often without understanding why, a result that leads to an overestimate of the students' capability, in a similar way to multiple-choice items. Identifying a "goldilocks" prompt, not too prescriptive, not too vague, often requires a multiple trial-and-error process.

8.3.1 The Role of Feedback Up to now, we have not addressed the role of feedback in assessment. Assessments can generally be divided into formative and summative. In the past, particularly at the university level, most assessment has been summative where the purpose of the assessment is to sort students, assign a grade, or allow them to pass through a gateway to further study. However, if we are to teach, that is, to introduce students to and help them develop more expert-like thinking that involves more than right/wrong, it becomes important to use formative assessment tasks to provide feedback to both student and instructor. Feedback to the student can, if done appropriately, drive their learning. Models of such interaction include the Socratic dialogue and the peer-review and author revision process, where students (authors) are asked to articulate their assumptions and explain and justify their reasoning. Feedback to the instructor is just as important: it can provide evidence about where students "are" so that instruction can be tailored to address areas of confusion or omission, what students do not know and perhaps have not been taught, as well as what they need to know and consider as relevant in order to construct an adequate answer. It can also be an impetus for course and curriculum redesign. Formative assessment is a stepping-stone to subject mastery and improved instruction.

For formative assessment to be effective, we know that students must receive timely and appropriate feedback (Hattie & Timperley, 2007). However, it has also been shown that, for various reasons, many students do not actually use the feedback they receive (Winstone et al., 2017) (see Chap. 9 in this volume). Nicol and Mcfarlane-Dick (2006) point out that while higher education has moved toward constructivist pedagogies, a parallel shift in formative assessment and feedback has generally not occurred. That is, formative assessment and feedback are seen as the responsibility of the instructor, where the feedback is a transmission process from instructor to student. Providing information about whether students responses are right or wrong, and perhaps even their strengths and weaknesses, may not bring about the kinds of changes we imagine or hope for. For example, some feedback is complex and not easily decoded and translated by the student into effective action. Responding to, and learning from feedback ~~it may well~~ requires the same kinds of opportunities to actively construct and test the validity of the student's understanding. ~~meaning as learning itself~~. For feedback to be effective, both the giver and the receiver of the feedback must understand the purpose and must engage with the feedback in a meaningful way. This often involves an extended dialogue—which of course means more time and effort on the part of both instructor and student: we will discuss approaches to this later.

Winstone et al. (2017) interviewed students and found that there were a number of barriers that deterred the use of feedback. For example, some feedback was too technical and could not readily be implemented by the student. The most common problems with feedback, however, had to do with grades. Students tend to ignore feedback if they have received a good (enough) grade; conversely if their grade was poor, many students did not use the feedback provided. Prior unrealistic expectations about the time and effort it would take to improve led to a kind of helplessness and the feeling that using feedback would be pointless. Often instructors can contribute in negative ways by conveying their eagerness (need) to move on, to "cover" the course content—reinforcing the implicit message that individual student learning is of secondary

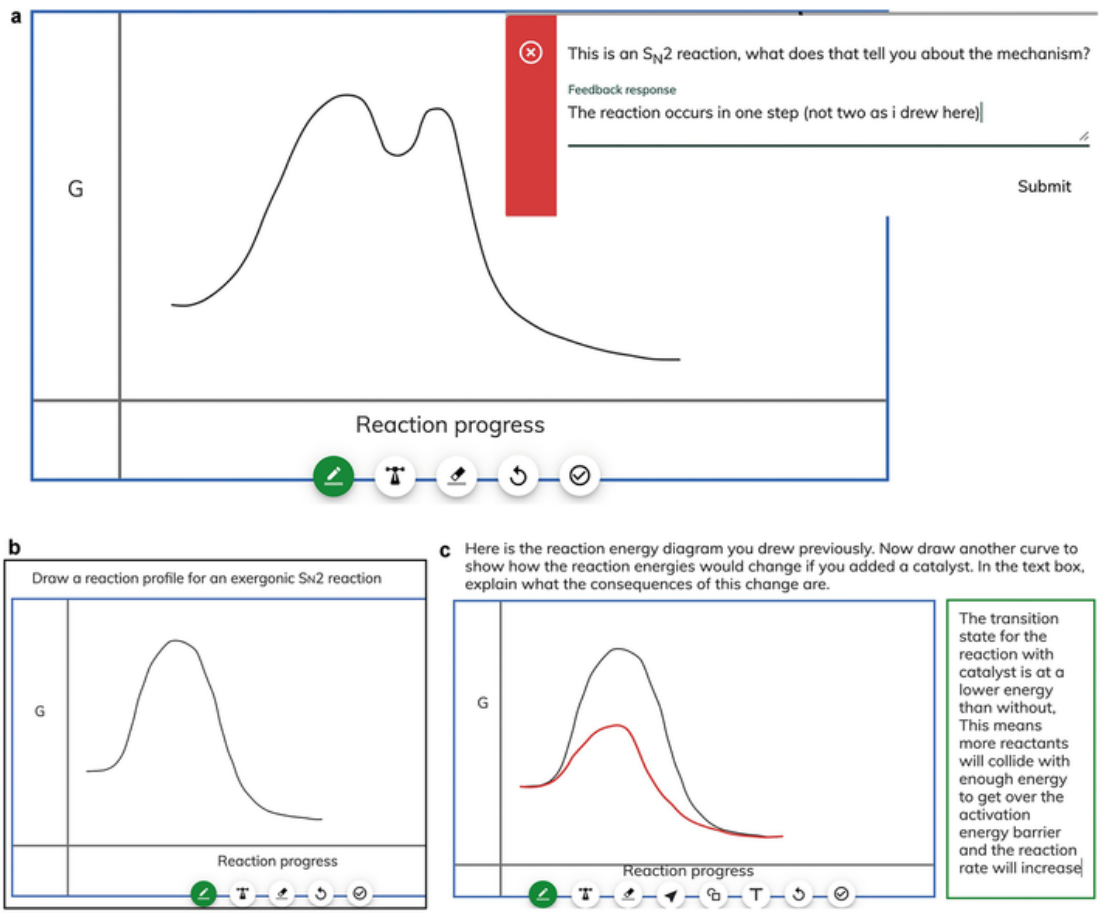
importance. It is suggested that for students to develop the agency and self-regulation to effectively use feedback, it should not be provided in the context of a graded assignment. Additionally, students should have the opportunity and time needed to engage with the feedback and consider (by themselves, or with others) its meaning and how it can be used.

8.3.2 Instructional Technology, Formative Assessment, and Feedback Most learning management systems provide the ability for multiple-choice assessments to provide feedback about right and wrong answers, as well as feedback of various types for incorrect responses. More sophisticated versions are found in "cognitive tutors" (D'mello & Graesser, 2013) but tend to be focused on mathematics, usually algebra (Pane et al., 2014), and the development of automaticity for rote cognitive skills or for training of instrument use. Such adaptive approaches have also found use to help medical students interpret diagnostic tests (Romito et al., 2016). However, if we want to support students, and provide feedback for more complex (constructivist) tasks, the technological solution must provide the capability to respond to students drawing and writing.

As noted earlier, beSocratic allows instructors to provide contextual responses to some kinds of simple diagrams, drawings, and graphs. For example, we might ask a student to draw an energy profile for a reaction, draw curved arrows to indicate an organic reaction mechanism, indicate the direction of energy transfer in a system, or predict the behavior over time of a gene expression network. beSocratic uses a system of rules inputted by the author/instructor, and feedback can be designed based on the context. Figure 8.4a shows a sample activity in which a student is asked to provide evidence that they know what kind of reaction is being discussed and what the energy change during the reaction is.

Fig. 8.4

(a) Feedback provided to students after drawing an incorrect energy diagram. (b) The correct diagram. (c) A new task showing how the student draws a new curve relative to the original, correct energy diagram



The student draws a response, which in this case is incorrect (there should be only one maximum), and receives feedback that they must respond to and then attempts to redraw the diagram. The feedback can be in increasingly directed tiers, if desired. When the student has successfully drawn the initial diagram (Fig. 8.4b), it can be transferred to the next task, where the student must respond to a new question (Fig. 8.4c).

This example shows some of the ways in which beSocratic can be used, by responding to graphical input, bringing forward prior drawings for comparison to use in a new task, and how written and drawn responses may be combined to provide a fuller response and more evidence about what the student knows and can do. We have developed over 100 beSocratic activities with these capabilities that are in regular use in the context of reimagined general and organic chemistry and biology courses (Cooper et al., 2019; Cooper & Klymkowsky, 2013; Klymkowsky et al., 2016). We have published a range of studies where we have compared students from transformed courses that use beSocratic activities as formative assessments to comparable cohorts of students who are enrolled in more traditional courses that use more conventional assessment systems. These studies reveal that students in transformed courses are better able to draw Lewis structures (Cooper et al., 2012), predict properties (Underwood et al., 2015), explain acid-base reactions (Cooper et al., 2016; Crandell et al., 2019), and nucleophilic substitutions (Crandell et al., 2020), among others.

For us, the next step is to integrate feedback into the written components of student explanations. We have already shown that we can train a machine grader to recognize differing levels of sophistication in student responses (Noyes et al., 2020), but what we cannot do yet is provide students with appropriate, useful discipline-specific feedback in a timely manner. Our goal over the next few years is to integrate beSocratic with natural language processing and machine learning systems to provide such feedback (Foltz et al., 2000).

We would be remiss if we did not acknowledge that there are a large range of adaptive technologies available for instruction and assessment, and these are discussed in detail in Chap. 5 in this volume.

8.4 The Future of Technology-Based Assessment By now we hope to have convinced you that the future of assessment lies not in more complex types of forced-choice instruments (Klymkowsky & Garvin-Doxas, 2020), but rather in well-designed open response tasks that require both written and drawn elements that can be automatically recognized, responded to, and graded (if necessary). Indeed, many of the elements necessary to make this a reality are already in place; we can recognize and respond to a subset of student drawn elements, and we can train systems to identify different levels of written responses, based on original human coding. Systems involving chatbots and the more sophisticated "smart conversational AI" are commonplace, often taking the place of human customer service. Certainly, these systems can be frustrating to deal with on occasion (Powell, 2019), but it is not a stretch to imagine that in the future students will interact with AI systems by drawing, writing, and speaking, with the AI providing accurate and relevant contextual feedback, scaffolded support, and challenging (Socratic) questions or grading students' responses. We envision a future in which examinations are no longer necessary, where concerns about cheating can be minimized (see Chap. 10 in this volume), where continuous formative tasks provide students with the opportunities needed to develop more expert-like levels of thinking, and where the evidence elicited by them is convincing and robust.

References

Becker, N. M., Noyes, K., & Cooper, M. M. (2016). Characterizing students' mechanistic reasoning about London dispersion forces. *Journal of Chemical Education*. https://doi.org/10.1021/acs.jchemed.6b00298

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33–63.

Bryfczynski, S. P. (2012). *BeSocratic: An intelligent tutoring system for the recognition, evaluation, and analysis of free-form student input* (UMI No. 3550201). Doctoral dissertation. Clemson University.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604–616.

Cooper, M. M., & Klymkowsky, M. W. (2013). Chemistry, life, the universe and everything: A new approach to general chemistry, and a model for curriculum reform. *Journal of Chemical Education, 90*, 1116–1122. https://doi.org/10.1020/ed300456y

Cooper, M. M., Kouyoumdjian, H., & Underwood, S. M. (2016). Investigating students' reasoning about acid–base reactions. *Journal of Chemical Education, 93*(10), 1703–1712. https://doi.org/10.1021/acs.jchemed.6b00417

Cooper, M. M., Stowe, R. L., Crandell, O. M., & Klymkowsky, M. W. (2019). Organic chemistry, life, the universe and everything (OCLUE): A transformed organic chemistry curriculum. *Journal of Chemical Education*. https://doi.org/10.1021/acs.jchemed.9b00401

Cooper, M. M., Underwood, S. M., Bryfczynski, S. P., & Klymkowsky, M. W. (2014). A short history of the use of technology to model and analyze student data for teaching and research. In R. S. Cole & D. Bunce (Eds.), *Tools of chemistry education research* (pp. 219–239). American Chemical Society.

Cooper, M. M., Underwood, S. M., Hilley, C. Z., & Klymkowsky, M. W. (2012). Development and assessment of a molecular structure and properties learning progression. *Journal of Chemical Education, 89*(11), 1351–1357. https://doi.org/10.1021/ed300083a

Cooper, M. M., Williams, L. C., & Underwood, S. M. (2015). Student understanding of intermolecular forces: A multimodal study. *Journal of Chemical Education, 92*(8), 1288–1298. https://doi.org/10.1021/acs.jchemed.5b00169

Crandell, O. M., Kouyoumdjian, H., Underwood, S. M., & Cooper, M. M. (2019). Reasoning about reactions in organic chemistry: Starting it in general chemistry. *Journal of Chemical Education, 96*(2), 213–226. https://doi.org/10.1021/acs.jchemed.8b00784

Crandell, O. M., Lockhart, M. A., & Cooper, M. M. (2020). Arrows on the page are not a good gauge: Evidence for the importance of causal mechanistic explanations about nucleophilic substitution in organic chemistry. *Journal of Chemical Education, 97*(2), 313–327. https://doi.org/10.1021/acs.jchemed.9b00815

D'mello, S., & Graesser, A. (2013). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*(4), 1–39.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*(2), 111–127.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gagne, R. M. (1965). *The conditions of learning*. Holt Rinehart & Winston.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74.

Hammer, D. (2000). Student resources for learning introductory physics. *American Journal of Physics, 68*, S52–S59.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141–158. https://doi.org/10.1119/1.2343497

Klymkowsky, M. W., & Garvin-Doxas, K. (2020). Concept inventories: Design, application, uses, limitations, and next steps. In *Active learning in college science* (pp. 775–790). Springer.

Klymkowsky, M. W., Rentsch, J. D., Begovic, E., & Cooper, M. M. (2016). The design and transformation of biofundamentals: A nonsurvey introductory evolutionary and molecular biology course. *CBE Life Sciences Education, 15*(4). https://doi.org/10.1187/cbe.16-03-0142

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*(2), 115–136. https://doi.org/10.1080/08957347.2011.554604

Miller, C. M., & Parlett, M. (1974). *Up to the mark: A study of the examination game*. Society for Research into Higher Education.

Mislevy, R. J., & Riconscente, M. M. (2011). Evidence-centered assessment design. In S. Downing & T. Haladyna (Eds.), *Handbook of test development*. Erlbaum.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. https://doi.org/10.17226/24783.

National Research Council. (1999). *How people learn: Brain, mind, experience, and school*. National Academies Press.

National Research Council. (2001). In J. W. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), *Knowing what students know: The science and design of educational assessment*. National Academies Press.

National Research Council. (2012). In S. R. Singer, N. R. Nielson, & H. A. Schweingruber (Eds.), *Discipline-based education r esearch: Understanding and improving learning in undergraduate science and engineering*. National Academies Press.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Noyes, K., & Cooper, M. M. (2019). Investigating student understanding of london dispersion forces: A longitudinal study. *Jo urnal of Chemical Education*. https://doi.org/10.1021/acs.jchemed.9b00455

Noyes, K., McKay, R. L., Neumann, M., Haudek, K. C., & Cooper, M. M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education, 97*(11), 3923–3936. https://doi.org/10.1021/acs.jchemed.0c00445

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

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a the ory of conceptual change. *Science Education, 66*(2), 211–227.

Powell, J. (2019). Trust me, I'm a chatbot: How artificial intelligence in health care fails the Turing test. *Journal of Medical Int ernet Research, 21*(10), e16222.

Romito, B. T., Krasne, S., Kellman, P. J., & Dhillon, A. (2016). The impact of a perceptual and adaptive learning module on tra nsoesophageal echocardiography interpretation by anaesthesiology residents. *BJA: British Journal of Anaesthesia, 117*(4), 477–481.

Schmidt, H.-J., Kaufmann, B., & Treagust, D. F. (2009). Students' understanding of boiling points and intermolecular forces. *Chemistry Education Research and Practice, 10*, 265–272. https://doi.org/10.1039/B920829C

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14. https://doi.org/10.3102/0013189X029007004

Skinner, B. F. (1954). The science of learning and the art of teaching. *Cambridge, Mass, USA, 99*, 113.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Focus article: Implications of research on children's learning fo r standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: I nterdisciplinary Research & Perspective, 4*(1–2), 1–98.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *Internatio nal Journal of Science Education, 10*, 159–169. https://doi.org/10.1080/0950069880100204

Underwood, S. M., Reyes-Gastelum, D., & Cooper, M. M. (2015). Answering the questions of whether and when student lear ning occurs: Using discrete-time survival analysis to investigate how college chemistry students' understanding of structur e-property relationships evolves. *Science Education, 99*(6), 1055–1072. https://doi.org/10.1002/sce.21183

Urban-Lurain, M., Cooper, M. M., Haudek, K. C., Kaplan, J. J., Knight, J. K., Lemons, P. P., Lira, C. T., Merrill, J. E., Nehm, R. H., Prevost, L. B., Smith, M. K., & Sydlik, M. (2015). Expanding a national network for Automated Analysis of Constructed Response assessments to reveal student thinking in STEM. *Computers in Education Journal, 6*(1), 65–81.

Williams, L. C., Underwood, S. M., Klymkowsky, M. W., & Cooper, M. M. (2015). Are noncovalent interactions an Achilles heel in chemistry education? A comparison of instructional approaches. *Journal of Chemical Education, 92*, 1979–1987. https://doi.org/10.1021/acs.jchemed.5b00619

Winstone, N. E., Nash, R. A., Rowntree, J., & Parker, M. (2017). 'It'd be useful, but I wouldn't use it': Barriers to university students' feedback seeking and recipience. *Studies in Higher Education, 42*(11), 2026–2041. https://doi.org/10.1080/03075079.2015.1130032

1 Reverse Dunning-Kruger effects and science education