Redesigning testing: operationalizing the new science of learning

Zachary Stein
Developmental Testing Service, LCC
Harvard University Graduate School of Education

Theo Dawson
Developmental Testing Service, LCC

Kurt W. Fischer
Harvard University Graduate School of Education

A revised and expanded version of this paper appeared in *The new science of learning: computers, cognition and collaboration in education*, Springer Press

Abstract: Complex standardized testing infrastructures have come to shape most educational systems. With so many people taking so many tests, we must seriously begin to ask, *what are we measuring?* and *what is worth measuring?* This paper presents the work of a research group that has begun to use the latest in computer technology and learning science to build tests that are both standardized and formative, grounded in research about learning, and richly educative. Fischer's *Dynamic Skill Theory* provides a framework for modeling the diverse learning sequences and developmental pathways that characterize how real individuals in real-world contexts learn and develop. Dawson's *Lectical™ Assessment System* is a psychometrically validated domain-general developmental assessment system. These two sophisticated outgrowths of contemporary learning science are being employed in an effort to design a new kind of testing infrastructure, an effort known as the *DiscoTest™ Initiative*. In this paper we describe these efforts and explore how these specific advances in research and design will change the practice of testing, beyond using standardized tests as mere sorting mechanism and toward the use of tests as educative aids.

List of figures:

Introduction: testing testing

Every year, across the globe, tens of millions of children, adolescents, and adults from all walks of life take tests. In the United States students may take anywhere from six to twenty standardized tests on their way from kindergarten to college, not counting the numerous summative and formative assessments employed by teachers. Imagine a high school student in Massachusetts who sits down to take a standardized test that will ultimately determine both her chances of graduation and the standing of her school. She participates in a large, complex, and polycentric educational testing infrastructure that transcends local, state, national, and international borders. At all points there are overlapping networks of connections with industry, government, and research. This international testing infrastructure is an unprecedented state of affairs, representing both a vast and incomparable example of "applied psychology" and a crucial force shaping educational systems. The goal of this paper is to begin to reflect on this state of affairs, bring key issues to light, and report on specific avenues of research and design for building a new type of educational testing infrastructure that will bring greater benefit to greater numbers by serving more diverse purposes and populations.

The current state of educational testing is the outcome of a complex history of educational research, practice, and policy. In the first section we draw out key themes from this history, framing the discussions to follow. Tests and assessments have always been a necessary aspect of most educational situations—being part of the conversation between teacher, student, and curriculum. As broad social and cultural trends towards mass schooling emerged, educational practice began to assimilate the outputs of a newly professionalized psychology, fostering the development of a specific type of testing infrastructure. This late-modern approach to testing is relatively isolated from research about learning, emphasizing the use of tests as sorting mechanisms, and neglecting the use of tests as educative aids. More recently, in the United States in particular, this approach to testing has been wedded to sweeping educational policies mandating specific high-stakes uses, which has put testing at the

center of many debates about schooling. This heightened awareness leads us to suggest that *now is the time* to ask foundational questions about what today's tests measure and how they are used.

In the second section, we begin a response to these foundational questions, suggesting that a testing infrastructure based on research into the nature of learning will be better able to meet the challenges facing educational systems in the 21st century. Arguments from a variety sources propose that the new science of learning should be at the heart of efforts to re-design testing infrastructures. Moreover, given the rapidly changing conditions to which educational institutions must respond, the values that shape test reform efforts should transcend outdated dichotomies about the function of testing and the purposes of education—moving beyond unproductive either/or commitments: either tests as sorting mechanisms or tests as educative aids; either tests of competencies or tests of content; either tests to train the work force or tests to foster reflective citizens. Tests should be based on research about how students learn and guided by explicit commitments to re-shaping schools in positive new directions.

In the third and fourth sections, we outline our approach to test development, wherein new computer-based tools are wedded to advances in psychometrics and cognitive developmental psychology, thus bringing the new science of learning to bear in the design of a broad and flexible testing infrastructure that is both standardized and formative. For decades research in cognitive development has focused on the diverse *learning sequences* that characterize the acquisition of knowledge, capabilities, or skills. Recently, in the wake of Fischer's *Dynamic Skill Theory*, a common or general scale has been built, which can be used to research and understand development and learning along an almost endless variety of different learning sequences. The *Lectical™ Assessment System* is a psychometrically refined measure of this general scale, allowing for reliable and valid assessments of student performance and the concomitant construction of empirically grounded learning sequences. The *DiscoTest™ Initiative* embodies our general approach to test design, which combines this

approach to researching and measuring learning—wherein diverse learning sequences can be understood in terms of a common scale—with advances in computer-based tools. The result is a radically new approach to testing, an approach that could form the foundation for *a mass customized testing infrastructure* that provides all the benefits of embedded, formative tests, with the kind objectivity and validity that are desirable in standardized tests. Moreover, as discussed in the conclusion, this new approach to test design re-frames what is considered possible and preferable for the future of the testing infrastructures that shape educational institutions.

Historical preamble: the broad function of testing and the birth of a specific testing industry

Education, broadly construed, serves a basic function in fostering crucial skills and dispositions in younger generations, thus enabling the continuity and reconstruction of social structures and cultural traditions (Dewey, 1916; Habermas, 1984). Comparative psychology suggests that sustained and explicit teaching and learning are unique to our species. While some other species pass on acquired techniques, some argue that no other species fundamentally depends on mechanisms of cultural transmission to foster the maturity of its members (Tomasello, 1999). In a definitive way, to be human is to be educated. Importantly, educational processes of almost any type depend upon assessments, or tests.[1]

Tests, broadly construed, serve a basic educational function. They are a necessary part of a dialogue between the student, the teacher, and what is being taught. For as long as *homo sapiens sapiens* has existed there have always been students and teachers because there have always been things to be taught. Thus, there have always been tests. Even before the invention

[1] Throughout this paper we use the terms *test* and *assessment* somewhat interchangeably, more commonly using the former. Both terms are rich with connotations and there are liabilities accompanying the use of either. We feel that *testing* better conveys a formalized educational process, whereas *assessment* is a more general and ambiguous term, which includes research instruments and various non-educative measurements of capability. We realize that our usage of these terms cuts against the grain of some aspects of common usage, but we desire to redeem *testing* from its status as a term of derision.

of schooling, informal and formal tests of all kinds were used for educational purposes, from the passing on of food procuring practices and culture specific skills to apprentice workshops and religious training (Cremin, 1970). In order for teachers to provide instruction or guidance they must understand what the student has understood so far. How else can the teacher know what the student needs to learn next? Testing is one primary way that the intergenerational interactions constituting cultural transmission become explicitly and reflectively educative. Thus the use of tests to "measure" student understanding has a long history. Yet, as discussed below, questions of what is worth measuring have not figured prominently in modern test design.

After the invention of schooling, formal testing itself became an explicit component of educational systems of various sizes and types. As testing became explicit its uses became more varied. Classically, public debates and oral exams came to supplement ongoing educative assessments, serving to determine if students had learned sufficiently to assume the roles in society they were being trained to fill (e.g., the priesthood). Proficiency in reading and writing became a focus of testing as some elite segments of the population came to value and require literacy. Thus, early on, beyond serving as an educative aid, formal testing infrastructures came to function as mechanisms serving social goals and perpetuating specific social structures. The use of tests as sorting mechanisms has a long history, and the privileging of this usage is a key theme in modern test design.

The birth of democracies fueled ambitions for large-scale public educational systems, and the emergence of these institutions coincided with the emergence of psychology as a discipline (Karier, 1986). This is a coincidence of no small import in the history of testing. Around the turn of the 20th century, psychologically informed testing procedures proliferated, spawning the field of psychometrics and the preliminary use of intelligence testing to administer mass schooling in France (Lagemann, 2000). Knowledge of IQ-testing broke into public awareness during the First World War, as the US Armed Forces pioneered large-scale

administrative applications of psychological testing—applications that were immediately adopted for educational use (Sokal, 1990). Despite the lamentable misuses of IQ-testing due to its ties to eugenics (Gould, 1981), by the end of the Second World War the Educational Testing Service had been founded, and our contemporary standardized educational testing infrastructure was beginning to take shape (Lemann, 1999).

The contemporary educational scene in most industrialized countries is dominated by a specific type of standardized testing infrastructure (Hursh, 2008; National Research Council, 2001). This is an infrastructure that has been shaped by the demands of rapidly growing public education systems with unprecedented influxes of students being educated for unprecedented amounts of time. Today's tests were built during radical social transformations that brought to light dire inequalities of educational opportunity and accomplishment. And, for the most part, the recent architects of our testing infrastructure have been adamant proponents of the fair distribution of educational opportunities and well aware of the important social function to be performed by the tests they designed (Lemann, 1999; Sokal, 1990).

However, our current testing infrastructure has been shaped by specific psychometric techniques and psychological commitments, criticized by one authority as "the application of 20th century statistics to 19th century psychology" (Mislevy, 1993, p. 19). Moreover, this approach to psychological testing has always neglected the educative function of tests and emphasized their use as sorting mechanisms for allotting future educational opportunities and conferring credentials (Chapman, 1988). The use of tests as sorting mechanisms allows the testing infrastructure to serve a broad public function in overseeing social role allocation. Thus, what now exists is an infrastructure built and run by private companies but serving a public function (Lemann, 1999). This has led to concerns about the existence of a standardized testing industrial complex and other socio-political criticism of the testing industry, from inflammatory exposés (Nairn, 1980) to more carefully reasoned calls for reform (Hursh, 2008).

In the United States this testing industry has been coupled to legislative injunctions resulting in the near universal use of high-stakes tests, which serve as both accountability measures for schools and graduation requirements for students (Hess & Petrilli, 2006; NRC, 1999). This nationally mandated use of a specific form of testing in K-12 education represents a radical departure from prior US educational policy, which had traditionally left control of test use and design up to state and local officials. These recent developments have increasingly brought testing into the center of national debates about education. In particular, the Obama administration has drawn attention to the liabilities of the contemporary testing infrastructure—in both its specific details and broad impacts (Obama, 2008; White House, 2009).

Teaching to the tests as they currently exits means preparing students for life as if it were a set of multiple-choice questions. It would seem that the time is ripe for seriously questioning the foundations of our testing infrastructure, asking a series of very basic questions: *What is being measured with today's tests? What should be measured? How are today's tests being used? How should they be used?*

Reflecting on testing: the need for a theory of learning and clarity about values

Although it may not be apparent at first, questions about what we are measuring with tests and how they should be used hinge upon the way we conceive the nature of *learning* (NRC, 2001). The criticism of contemporary tests as the application of advanced statistics in service of simplistic psychology is to the point. In order to use tests effectively and knowledgeably we need to understand the meaning of the score a student receives. Does a score mean something about a capability or trait processed by the student, or does it simply let us know how the student performs on a specific set of questions in relation to group averages? The latter claim—given that it remains strictly descriptive, positing no explanation for the score (e.g., IQ) or prescription for changing it—does not entail beliefs about specific psychological constructs; the former does.

Claiming that a test score reflects an underlying capability or trait—be it an aptitude, skill, or disposition—entails certain views about these psychological phenomena. More specifically, using such a claim about the meaning of a test score to guide actions, such as doling out remediation or rewards, entails some theory of *learning*. Imagine again the high school student from Massachusetts taking a standardized test. The use to which her scores will be put—determining graduation eligibility and school standing or quality—imply that the capabilities being assessed are the result of her individual effort and the school environment. That is, they assume a theory—however implicit—about how the capability being measured changes over time as a result of certain factors. Roughly speaking, theories about how psychological phenomena undergo change are theories of learning or development (Reisberg, 2001). Different theories of learning will give the same test score different meanings, and different theories of learning result in different forms of test design (NRC, 2001; See figure 1). Moreover, a test built without an explicit theory of learning—as many tests are—can serve only very limited functions.

For example, most standardized tests, like the one taken by the Massachusetts student, can serve *only* as sorting mechanisms because they are built without reference to an explicit theory of learning. No doubt, they are reliably and objectively measuring *something*, but it is unclear how this "something" relates to the learning process. Thus, tests not carefully wedded to a theory of learning can be used to classify students and schools, to sort them, but such tests cannot be used as educative aids, because they provide no insight into the learning process *per se*. On the other hand, building tests around an explicit theory of learning increases the range of functions the test can serve, e.g., allowing for insight into what a student has learned and could most benefit from learning next.

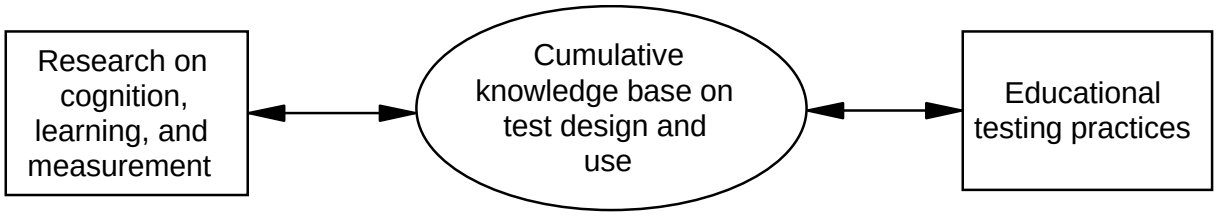


Figure 1: Connections between research, test design, and practice. Adapted from National Research Council (2001, p. 295).

Contemporary theories of learning have become increasingly sophisticated due to advances in cognitive science and neuroscience. Likewise, advances in psychometrics have made it possible to reliably and validly measure a wider range of dynamic and meaningful constructs. As alluded to in the historical section above, the contemporary testing infrastructure, and the uses to which it is put, reflect these advances only in a very limited way, if at all. The rest of this paper is devoted to outlining one approach that pulls together advances in the new science of learning and psychometrics to re-tool test design and educational practice.

Fischer's *Dynamic Skill Theory* (Fischer & Bidell, 2006) provides a comprehensive and empirically grounded approach to human development and learning. This approach has fostered methods for building empirically grounded learning sequences in a variety of domains, which can be aligned along a common underlying developmental dimension. This underlying dimension has been operationalized as a psychometrically refined metric, known as *The Lectical Assessment System* (the LAS: Dawson, 2008), which has been used to build unique and richly educative tests that are both standardized and customizable to different curricular frameworks (Dawson & Stein, 2008). The *Discotest Initiative* is the name given to our efforts in this direction. Below we discuss this approach to redesigning standardized testing infrastructures based on the new sciences of learning; we explain why it should be seen as a valuable alternative to the infrastructures currently in place.

However, the value of different testing infrastructures should not be determined solely on the basis of the manner in which they wed research about learning with test design and

educational practices. Decisions must be made about the general shape of the educational system and the role that ought to be played by even the best-designed tests. These decisions are fundamentally evaluative. They touch on some of our broadest goals and commitments about how schooling fits into a shared vision of the good life and just society. Typically, a set of classic dichotomies have been used to frame this discussion: Should tests be used to sort individuals and make high-stakes decisions or should tests be embedded in the curricula as educative aids? Should tests assess broad competencies or specific knowledge? Should tests help us in training the workforce or in fostering critically minded citizens?

Below, we will show that these are false dichotomies. Advances in test design allow for a reevaluation of what is generally considered as possible and preferable for mass-education and its testing infrastructure. After discussing the approach we have adopted for designing tests based on the new science of learning, we will return to some of the broad evaluative questions and discuss how we understand the implications of these innovations.

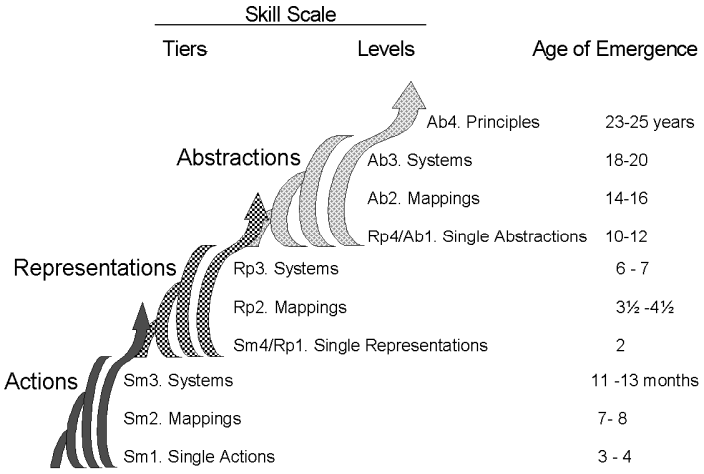
Advances in developmental science and the birth of the DiscoTest initiative

James Mark Baldwin (1906) was the first psychologist to offer a complex view of human development in which a variety of different *learning sequences* unfold across qualitatively distinct developmental levels. This set an important agenda for development science, wherein a *learning sequence* is defined as an empirically grounded reconstruction of the levels or phases undergone during the acquisition of a specific capability, concept, or understanding. Decades after Baldwin, Heinz Werner (1957) and Jean Piaget (1932) would also offer theories of human development in which learning sequences figured prominently. Eventually, Kohlberg (1984) would build learning sequences in the moral domain, King and Kitchener (1994) in reflective judgment, Case (1992) in several knowledge areas, Watson and Fischer (1980) for social roles, and Siegler (1981) in mathematics, with many others following suit. For over a century researchers have been creating new methods and building empirically grounded models of

specific learning sequences in a wide variety of domains. This general approach to researching development and learning continues to produce knowledge, with an increasing focus on individual differences and educational implications (Stein, 2009).

As a part of this tradition, Fischer's *Dynamic Skill Theory* (Fischer, 1980; Fischer & Bidell, 2006) has added a generative set of methods and concepts useful for researching and modeling learning sequences. First outlined in the 1980s, the General Skill Scale (Figure 2) is the backbone of the general approach. The Skill Scale is a model of the basic structural transformations characteristic of skill development and has been empirically refined in light of decades of research. Importantly, in this context the term *skill* should be taken in a very general sense, as the basic or generic unit of psychological process. All skills are richly multidimensional, intrinsically involving emotion, cognition, context, and social support. Skills are built actively and dynamically by individuals in specific contexts and they are built hierarchically, with more complex ones transcending but including less complex ones. As individuals build unique skills in different domains, learning sequences unfold across the different tiers and levels: *actions* lay the groundwork for concrete *representations*, which serve eventually as the basis for the construction of *abstractions*. Within each tier, there is a series of levels, as the basic skill-type (action, representation, or abstraction) is coordinated into increasingly complex forms of organized behavior.

Figure 2. Developmental Scale for Skill Levels and Tiers



Thus the Skill Scale is a model of the developmental dimension underlying the learning sequences that comprise development in almost all domains. This model has been used to guide researchers in their analysis of behaviors and performances, informing the dissection of the skill structure and thus locating performances along the developmental continuum specified by the general Skill Scale—a technique known as Skill Analysis. Analyzing the structure of diverse performances as they unfold over time in a given domain allows for the inductive reconstruction of learning sequences in that domain. This general technique has been used in a variety of domains, including mathematics (Fischer, Hand, & Russell, 1984), reflective judgment (Kitchener & Fischer, 1990), and self-in relationship (Fischer & Kennedy, 1997). Research informed by the Skill Scale has been paralleled by research involving comparable frameworks (Case, 1992; Commons, Trudeau, Stein, Richards, & Krause, 1998), which has reinforced the idea that the Skill Scale represents an important underlying developmental dimension.

Importantly, this research tradition has focused in large part on the diversity and dynamism of human development. Thus, the emerging consensus regarding a common scale

should not be seen as a re-working of the simple ladder-like, growth-to-goodness models of development offered in the 1960s and 70s. Instead, learning sequences are understood as sets of diverse pathways, which individuals traverse in unique ways—often towards common goals (Fischer & Bidell, 2006). Of course, some researchers have constructed learning sequences to serve as ideal types—simplifying the dynamics of development into set of static level descriptions. These idealizations can be useful only insofar as they frame an understanding of how individual learners work, *in medias res*, to construct unique paths through this empirically grounded but ideally represented space. Thus, focusing on individual differences does not entail neglecting invariance. The general Skill Scale represents an important confluence of research concerning certain invariant processes underlying the diversity and variability of real human learning in context.

In the wake of this confluence of research, Dawson (2008) confirmed the existence of a developmental dimension underlying a wide variety of learning sequences by applying a set of psychometric techniques. This resulted in a refinement of the basic principles of Skill Analysis—along with other comparable developmental assessment systems—and the construction of the most psychometrically validated and reliable developmental assessment system to date, the *Lectical Assessment System* (LAS). The LAS has been used to systemize the construction of learning sequences out of both longitudinal and cross-sectional data sets (Dawson-Tunik, 2004). This process for building learning sequences involves a three step iterative method (described in detail elsewhere: Dawson-Tunik, 2004; Dawson & Stein, 2008).

Dynamic Skill Theory and the *Lectical Assessment System* represent fundamental advances both our understanding of learning and our methods for studying and measuring it. Importantly, the ability to build learning sequences about specific topics using a psychometrically sophisticated instrument allows for a radically new approach to test design. The DiscoTest Initiative seeks to build a new testing infrastructure around the basic advances provided by this broad approach to understanding and researching learning. Before getting into

the details of how DiscoTests are built and used, we will briefly explain the guiding insights and goals.

As explained above, tests should be built around actual research into how students learn (NRC, 2001). The systematic construction of focused learning sequences provides valuable insight into student learning processes by allowing for a general characterization of the range of possible conceptions—the steps along the way from less sophisticated to more sophisticated understandings. This allows for tests that can place any given student performance in relation to the range of possible performances, and thus gives insight into what the student currently understands and what the student is likely to benefit from learning next. These tests can be integrated with curricula that are also informed by empirically grounded learning sequences, which also can provide the basis for richly educative feedback. Moreover, because the learning sequences are built around a psychometrically refined general metric, a test that is scored on that metric also serves as a standardized measure of student performance. Thus, the goal of the DiscoTest initiative is to build standardized tests that can be customized to different curricula and built around empirical research into how students learn, providing both educative feedback and psychometrically reliable scores.

DiscoTest: building the computer based educational testing infrastructure of tomorrow

The DiscoTest Initiative is a non-profit, research-oriented effort to develop free, valid, reliable, standardized, and educative assessments of key skills and concepts. These assessments can easily be embedded in curricula and can be employed without extensive training to track student development in classrooms, schools, districts, or nations. Each DiscoTest (also called a *teaser*—short for *brain teaser*) is developed by a team of researchers, content experts, and teachers who have come together as peers to study the development of a "big idea" or core skill (e.g., the physics of energy, conservation of matter, algebraic thinking, the scientific method, reflective judgment, leadership, or ethical reasoning) and then use the results

to describe learning sequences for important concepts. These learning sequences are then used to inform curricula and construct low-inference scoring rubrics for one or more teasers.

The overarching objective of the DiscoTest Initiative is to contribute to the development of optimal learning environments by creating assessments that deliver the kind of educative feedback that learners need for optimal learning. Assessments of this kind determine where students are in their individual learning trajectories and provide feedback that points toward the next incremental step toward mastery. They function as standardized formative assessments.

Building tests with these qualities requires an entirely new approach—one that is discursive and iterative, bringing together educators, researchers, and domain experts as equal partners. The name “Disco” was chosen for this initiative because it is the Latin root of *discourse*. Coincidentally, it also evokes the image of joyful kinesthetic interactions with music, an image that sits well with the notion that learning is fun.

Naming the Disco initiative was the least of many challenges. Here are a few others: DiscoTests:

1. *must be grounded in solid empirical evidence about the ways in which students learn specific concepts and skills. (To accomplish this goal, Dawson developed a new set of research and test development methods.)*
2. must be composed of intriguing items that allow students to show how they think about what they have learned, rather than simply demonstrating that they can get a “right” answer.
3. must not waste students’ time. In other words, every interaction with a DiscoTest must be a useful learning experience, and all DiscoTests must funtion as an integral part of the curriculum.
4. must provide students, teachers, and parents with a record of learning in which each milestone is meaningfully connected to specific knowledge and skills.

5. must have a long shelf-life, which implies that (1) they are of enduring importance and that (2) it should be very difficult to cheat on them and (3) they should used in ways that make it seem pointless to cheat on them.
6. must provide data that researchers can use to continually refine our understanding of learning.

Although it is not possible to provide a detailed account of our approach to all of these challenges within the context of a brief chapter, in this section we show how several of them are addressed through the DiscoTest Initiative and the design of DiscoTests. First we describe the tests and how they are scored. Then we describe how we build them, and how they can be used.

Anatomy of a DiscoTest. Because teasers are designed to be tests in the sense described above, they must (1) provide students with an opportunity to engage in meaningful action on their knowledge and (2) offer useful feedback. These requirements forced us to rule out multiple choice items like the one shown in Figure 2. This item asks the student to select one of 5 possible responses. There is one right answer. The other answers are intended to represent common misconceptions held by students. A number of assumptions accompany items of this kind. For example, it is assumed that students who get an item right either (1) know the answer or (2) have made a good guess. Also, it is generally assumed that students who get an item right without guessing know more than students who get an item wrong. If these assumptions held, items of this kind might provide information that could inform accurate feedback—but the assumptions do not hold.

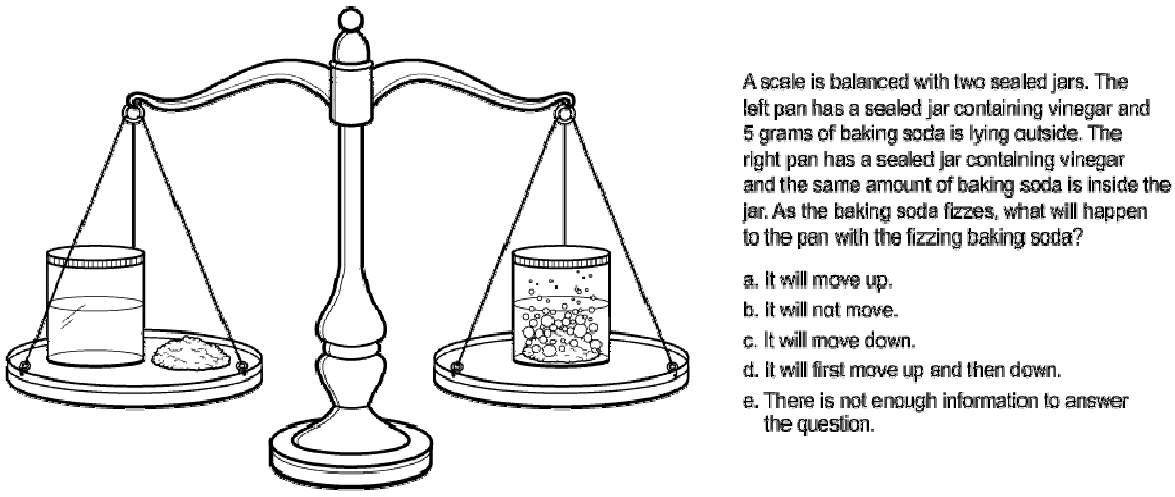
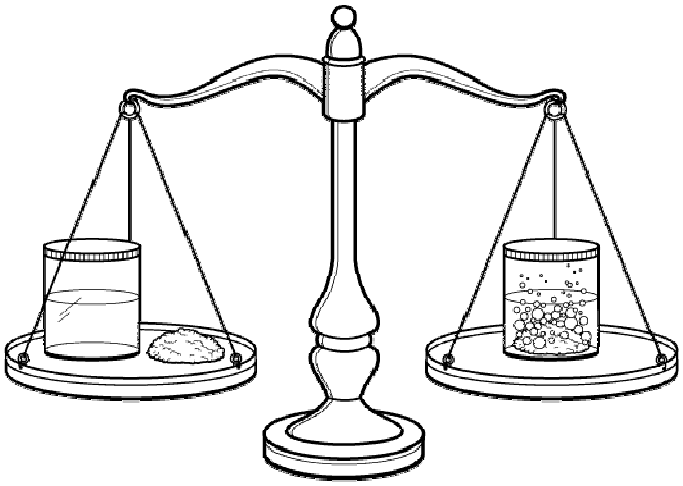


Figure 3: Multiple choice vinegar and baking soda item

Students who select the right answer (b) do so for several different reasons, many of which reflect partial knowledge or misunderstanding. Here are some of the explanations students have given for choosing the correct answer to the item in the example:

1. The right pan will not move because the amount of matter in a closed container remains the same no matter what chemical or physical changes take place. (textbook response, could be memorized)
2. The right pan will not move because a gas was formed but nothing was destroyed. (answer showing partial understanding)
3. The right pan will not move because not even bubbles can get out of a jar that is closed tight. (answer showing that the student believed the jar was closed really tight)
4. The right pan will not move because the gas doesn't have any weight. (answer showing partial understanding)
5. The right pan will not move because in a chemical reaction, atoms rearrange to make new substances, but none of them are destroyed. (answer showing greatest level of understanding)

This phenomenon, which has been described by numerous researchers (Sadler, 2000), strongly suggests that multiple choice items are unlikely to provide the kind of information required to inform educative feedback. Sadler has shown how multiple-choice tests can be used more effectively, but multiple choice items, no matter how well they are constructed, still limit the learning functions of an assessment. Consequently, DiscoTests are composed of items that are open-ended and require short essay responses consisting of judgments and justifications that not only show (1) what students know, but also (2) how they understand what they know and (3) how they can use their knowledge to deal with similar tasks and situations. The item in Figure 3, stripped of its multiple-choice options, can function as a DiscoTest item, as shown in Figure 4.



A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda is lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda is inside the jar. What will happen to the pan with the fizzing baking soda? Why? (Explain what happens in as much detail as possible, using what you have learned in class about problems of this kind.

"The pan with the baking soda inside the jar will move up because when vinegar and baking soda are mixed together, they make a gas that is lighter than air, so it goes up like a birthday baloon."

Figure 4: Open-ended vinegar and baking soda item

Teasers are commonly composed of 5 to 7 items of this kind, which provide information that cannot be provided by multiple-choice items. For example, they can help teachers answer questions like the following:

1. What concepts is this student working with?

2. How does she understand these concepts?
3. What is her line of reasoning?
4. How well does she explain her thinking?

In addition to helping students consolidate new knowledge, items of this kind provide an opportunity to hone essential life skills like reasoning and writing.

Scoring and reports. After students submit their responses to a set of teaser items, they are directed to a coding page on which they are asked to match their own responses to options in a series of "pull-down" menus. These menus function as *low inference* rubrics, and can be used effectively by students, teachers, and researchers. For example, the following choices are offered in one of the coding menus for the item shown in Figure 4:

- none of the codes are like statements made in this response
- the right pan will move because more (or less) stuff is in it
- the right pan will not move because nothing can get out
- the right pan will move because of something that happens when baking soda is mixed with vinegar
- the right pan will not move because nothing is created or destroyed
- the right pan will not move because new molecules have formed but the number of atoms is the same
- the right pan will not move because the mass of a closed system stays the same, no matter what kind of chemical reaction or state change takes place

Students simply choose the response that most closely matches their own. For students, coding is an important part of the learning process, because it allows them to reflect upon their own performances in light of a range of response options.

After their coding selections are submitted, students are redirected to a report that (1) portrays their score on Fischer's General Skill Scale (also known as the Lectical™ Scale), (2) describes what their performance suggests about their current level of understanding, and (3) provides suggestions for developmentally appropriate learning activities (Examples can be viewed at DiscoTest.org.). In addition to viewing individual reports, students who have taken an assessment on multiple occasions can track their own developmental progress by viewing a figure that shows how their thinking has developed over time.

Of course, teachers can also code teasers, which provides students with an expert perspective on their performances and helps build teachers' knowledge about how students learn particular concepts and skills. Teachers can view individual student reports and classroom profiles that show the distribution of scores relative to the General Skill Scale.[2]

Uses of DiscoTests. In the classroom, DiscoTest teasers can be used in a number of ways. They can be used to assess students' pre-instruction knowledge or stimulate student interest in a new subject-area. They also can be used during or following instruction to (1) test how well students understand new ideas, (2) stimulate class discussion, and (3) help students integrate new concepts into their existing knowledge. They can be taken by single students, groups of students, or the entire class, and can be scored by individuals or groups.

DiscoTests are ideal for informing parents of their children's progress in school, because they provide specific information about how well their child understands class material and what he or she is most likely to benefit from learning next. In fact, the learning suggestions that are included in each report often consist of activities that parents and children can do together. For example, a student struggling to grasp the basic concepts might be asked to watch an on-line video where backing soda and vinegar are combined, or to try this at home. A more advanced

[2] Sample assessments and reports can be found at http://discotest.org.

student might be pointed to a university website where the principles of conservation of matter are applied to real life engineering problems.

Finally, because student performances remain in the system indefinitely, entire schools or districts can follow the development of individual students over time, providing a high quality method for tracking student progress and evaluating curricula.[3] Moreover, cumulative evidence of growth over time and across several topics provides a basis for making judgments about student readiness to graduate or advance to a new level in their studies.

Building DiscoTests. Needless to say, constructing reliable and valid tests of the kind we have described requires methods that are not part of the toolkit of most test developers. The most fundamental hurdle has been working out a practical approach to describing learning sequences at a fine-enough grain to make them useful for diagnostic purposes. Our (still evolving) solution is a set of methods that make use of a developmental metric that can be used to measure progress along the General Skill Scale. This metric, the Lectical Assessment System (LAS), is a well-tested developmental scoring system that consistently produces scores that are reliable within about ¼ of a General Skill Level (known as a *phase*).[4] The LAS thus makes it possible to use a combination of longitudinal and cross-sectional data to construct accurate and detailed learning sequences.

Researchers and teachers use these methods to document the pathways through which students learn a specific skill or set of concepts, to build curricula, and to facilitate learning. The research process is iterative. We begin by designing a single interview instrument composed of a set of open-ended problems that can be used to study the thinking of children as young as 5. Then, making an effort to sample approximately 20 individuals performing in each phase

[3] Although DiscoTest does not collect identifying information for students, the system is set up so that each student is associated with a unique identifier.
[4] In a given classroom for grades 7 to 12, the range of student performances is likely to be 4 – 6 phases (1.5 skill levels).

represented in the age range from 5 to 20 or so (about 14 phases), we use the instrument to conduct probed, clinical interviews. These are independently (1) scored with the LAS to determine their developmental phase and (2) submitted to a comprehensive analysis of their content. When both analyses have been completed, analysts study the relation between the level of performance and their conceptual content, gradually constructing descriptions of understanding in each phase and connecting these to describe a detailed learning sequence. This process is described more thoroughly elsewhere (Dawson, 2002, 2003, 2004; Dawson & Gabrielian, 2003; Dawson, Xie, & Wilson, 2003; Dawson-Tunik, 2004).

The items from the interview instrument usually become items in the first version of a new DiscoTest. The newly described learning sequence and the interview data used to construct it provide the basis for initial versions of coding menus and student feedback. At this point, the new teaser can undergo a first round of testing. Two to three rounds of testing are required to refine coding menus, check the accuracy of the learning sequence, evaluate item functioning, and optimize reliability.

Unlike conventional tests, DiscoTests can be used indefinitely; students can take the same teaser several times without exhausting its potential to help them gain an increasingly sophisticated understanding of targeted concepts. This is because the items are deliberately constructed to be answerable at several different levels of sophistication. Moreover, because the primary role of DiscoTests is educative (and items don't have single correct answers) concerns about cheating are minimal. Furthermore, DiscoTests are both educative and standardized. All performances are placed on the same, domain independent, general scale. This makes it possible to compare learning across any range of subjects or contexts. Finally, DiscoTests double as data collection instruments. Eventually they will yield large longitudinal databases that will allow researchers to construct increasingly refined accounts of the pathways through which students learn important skills and concepts.

New tools foster new values: re-visioning education and testing

We have discussed the history of our contemporary testing infrastructure and explained the need for new approaches grounded in the science of learning. We have also provided an overview of one new approach that combines advances in basic research about learning with new techniques in psychometrics to build embedded formative assessments that are both standardized and richly educative. The DiscoTest initiative is engaged in building *a mass-customized testing infrastructure* wherein metrics that are informed by learning theory and research are used in the design of standardized tests that fit the needs of specific curricula and support learning. This way of designing tests has broad implications for what is considered possible and preferable for standardized testing infrastructures.

This approach to test design also allows us to transcend the dichotomies mentioned in the introduction and third section. First, by using a psychometrically refined developmental measure to research specific learning sequences and then using that same measure to assess student performance relative to the researched learning sequences, we transcend the dichotomy between educative and standardized assessment. Student performances are evaluated both in terms of where their performance is in relation to the Skill Scale and in relation to the learning sequence being assessed, a measurement that is simultaneously standardized and educationally relevant. Knowing where the student performs on the General Skill Scale provides all the benefits (and liabilities) of standardized testing (e.g., allowing for comparison between students, or between subject areas for the same student, or between groups of students). But knowing where a student performs relative to an empirically grounded learning sequence provides all the benefits of formative assessment; with rich information about what the student understands and could best benefit from learning next. With the right measures,

research approach, and curricula, we no longer must choose between tests that are standardized and tests that are educative aids.

Secondly, the dichotomy between testing for general capabilities and testing for specific content is rendered moot. Research reveals that skills that are conventionally thought of as general skills, such as those for abstract reasoning, critical thinking, or academic writing, unfold along unique pathways *in specific content areas*. Thus, just as a score on a DiscoTest is both standardized and educative, it is also indicative of a range of general skills as they are exercised in specific content areas. These skills are demonstrated when students explain their thinking in written responses.[5]

Thirdly, DiscoTests overcome the dichotomy between testing to prepare the workforce and testing to foster critically minded citizens. This classic dichotomy is an artifact of an earlier era, before post-industrial conditions characterized large segments of the world and information technologies created a networked poly-vocal global public sphere. Today, we face unique conditions that render traditional ideas about the nature of socialization and adult life obsolete. Patterns of parenting, friendship, work, marriage, and political involvement have shifted rapidly away from predictability towards diverse individualized pathways of socialization with multiple outcomes and divergent views of success (Arnett, 2004; Beck, 2001). This complexity and heterogeneity should be met by a flexible educational system capable of responding to the unique needs of an increasingly diverse population of students. Such a system, if it is to maintain rigor and efficacy, will need a testing infrastructure that is standardized and customizable, broad and flexible, one that integrates basic knowledge about learning into new contexts and applications—one that rewards good thinking rather than right answers. DiscoTests don't have right answers. They are designed to provide students with many

[5] It is important to note that leaning disabled students or students whose native language is not the language of instruction and assessment will need appropriate accommodations, probably along the lines of the principles of *Universal Design For Learning* (See: Rose & Meyer, 2002).

opportunities to apply their knowledge to the kinds of problems they will face in the real world—messy, open-ended problems without simple answers.

We have thus touched on the philosophical issues at the heart of testing reform. The contemporary testing infrastructure set constraints on pedagogical options and structures the distribution of opportunities and resources. Moreover, as many have noted (e.g., Hursh, 2008; Dewey, 1916), the reward systems of schools act as proxies for the general values of society. Tests teach students—both indirectly and directly—what is deemed valuable in the socio-cultural context they inhabit. Thus a new testing infrastructure will have wide ranging implications, from classroom practice to college admissions and beyond. Redesigning large scale testing infrastructures means, in part, re-casting how social values are operationalized. The possibilities for building fundamentally new types of tests, based on the new science of learning and human development, allows us to transcend narrow debates about the goals of schooling and to help people learn better.

Sources:

Arnett, J. (2004). *Emerging adulthood*. New York: Oxford University Press.

Baldwin, J. M. (1906). *Thought and things: A study in the development of meaning and thought or genetic logic* (Vol. 1-3). New York: Macmillan Co.

Beck, U. (2001). A life of one's own in a runaway world: Individualization, globalization and politics. In *Individualization* (pp. 22-30). London: SAGE Publications.

Case, R. (1992). *The minds staircase: Exploring the conceptual underpinnings of children's thought and knowledge*. Hillsdale, NJ: Lawrence Erlbaum.

Chapman, P., D. (1988). *Schools as sorters*. New York: New York University Press.

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review, 18*, 237-278.

Cremin, L. (1970). *American education: The colonial experience*. New York: Haper & Row.

Dawson, T. L. (2008, 11/1/04). The Lectical™ Assessment System. 1. Retrieved September, 2008, from http://www.lectica.info

Dawson, T. L. (2002). A comparison of three developmental stage scoring systems. *Journal of Applied Measurement, 3*, 146-189.

Dawson, T. L. (2003). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology, 164*, 335-364.

Dawson, T. L. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development, 11*, 71-85.

Dawson, T. L., & Gabrielian, S. (2003). Developing conceptions of authority and contract across the life-span: Two perspectives. *Developmental Review, 23*, 162-218.

Dawson, T. L., Xie, Y., & Wilson, M. (2003). Domain-general and domain-specific developmental assessments: Do they measure the same thing? *Cognitive Development, 18*, 61-78.

Dawson, T. L., & Stein, Z. (2008). Cycles of reseach and application in education: Learning pathways for energy conpcets. *Mind, Brain, and Education, 2*, 89-102.

Dawson-Tunik, T. L. (2004). "A good education is…" The development of evaluative thought across the life-span. *Genetic, Social, and General Psychology Monographs, 130*(1), 4-112.

Dewey, J. (1916). *Democracy and education*. New York: The Free Press.

Fischer, K. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477-531.

Fischer, K., & Bidell, T. (2006). Dynamic development of psychological structures in action and thought. In W. Damon & R.M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (6th ed. Vol. 1, pp. 313-399). New York: John Wiley & Sons.

Fischer, K. W., Hand, H. H., & Russel, S. (1984). The development of abstractions in adolescence and adulthood. In M. L. Commons, F. A. Richards & C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development* (pp. 43-73). New York: Praeger.

Fischer, K. W., & Kennedy, B. (1997). Tools for analyzing the many shapes of development: The case of self-in-relationships in Korea. In K. A. Renninger & E. Amsel (Eds.), *Processes of development* (pp. 117-152). Erlbaum: Mahwah, NJ.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Habermas, J. (1984). *The theory of communicative action: Reason and the rationalization of society* (T. McCarthy, Trans. Vol. 1). Boston: Beacon Press.

Hess, F., & Petrilli, M. (2006). *No child left behind*. New York: Peter Lang.

Hursh, D. (2008). *High-stakes testing and the decline of teaching and learning*. New York: Rowman & Littlefeild.

Karier, C. (1986). *The individual, society, and education: A history of American educational ideas* (Second ed.). Chicago: University of Illinois Press.

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults*. San Francisco: Jossey Bass.

Kitchener, K. S., & Fischer, K. W. (1990). A skill approach to the development of reflective thinking. *Contributions to Human Development, 21*, 48-62.

Kohlberg, L. (1984). *The psychology of moral development: The nature and validity of moral stages* (Vol. 2). San Francisco: Jossey Bass.

Lagemann, E. (2000). *An elusive science: The troubling history of educational research*. Chicago: University of Chicago Press.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus and Grioux.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory of a new generation of tests* (pp. . Hillsdale, NJ.: Erlbaum.

Nairn, A. (1980). *The reign of ETS: The corporation that makes up minds*: The Ralph Nader Report on the Educational Testing Service.

National Research Council on the Foundations of Assessment (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, D.C.: National Academy Press.

National Research Council on the Foundations of Assessment (2001). *Knowing what students know: The science and design of educational assessment*. Wahington, D.C.: National Academy Press.

Obama, B (2008) Speech to the 146th Annual Meeting and 87th Representative Assembly of the National Educational Association. Delivered July 5th, 2008.

Piaget, J. (1932). *The moral judgment of the child*. London: Routledge and Kegan Paul.

Reisberg, D. (2001). Learning. In R. Wilson & F. Keil, C. (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 460-461). Cambridge, MA.: MIT press.

Rose, D. & Meyer, A. (2002) *Teaching every student in the digital age: universal design for learning*. ASCD Books. Washington, DC.

Sadler, P. M. (2000). The relevance of multiple choice tests in assessing science understanding. In J. J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), *Assessing*

science understanding: A human constructivist view (pp. 249-278). San Diego, CA: Academic Press.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46*, 84.

Sokal, M. (Ed.). (1990). *Psychological testing in American society: 1890-1930*. New Brunswick: Rutgers University Press.

Stein, Z. (2009). Resetting the stage: Introduction to special sections on learning and development. *Mind, Brain, and Education, 3*(2).

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Watson, M. W., & Fischer, K. W. (1980). Development of social roles in elicited and spontaneous behavior during the preschool years. *Developmental Psychology, 16*, 484-494.

Werner, H. (1957). The concept of development from a comparative and organismic point of view. In D. B. Harris (Ed.), *The concept of development*. Minneapolis: University of Minnesota Press.

White House office of the press secretary (2009) Administration's statement on educational policy. Retrieved June, 2009 (http://www.whitehouse.gov/the_press_office/Fact-Sheet-Expanding-the-Promise-of-Education-in-America/)