Intuitive Test Theory

Robert J. Mislevy
University of Maryland

Henry I. Braun
Educational Testing Service

May 17, 2001

This paper is the text of an invited address by the first author at the Annual Dinner Meeting of the Princeton Association for Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE) Computer Society Chapters, in Kingston, NJ, May 22, 2003. Thanks to Brian Berenbach, Jim Matey, and Dennis Mancl for extending this opportunity, and thanks to Paul Holland for stimulating conversations on the topic. The preparation of the presentation was supported by the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education. The findings and opinions expressed in this report are those of the authors, and do not reflect the positions or policies of Educational Testing Service or the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U. S. Department of Education.

Abstract

This presentation discusses some “phenomenological primitives” (p-prims) about testing and assessment. “P-prim” is a term coined by the psychologist Andrea diSessa to understand people's incorrect notions about science, such as “heavy objects fall faster than light objects”. Three examples of test theory p-prims are "A test measures what it says at the top of the page", "Any two tests that measure the same thing can be made interchangeable, with a little equating magic", and "Multiple-choice questions only measure recall." P-prims like these underlie nonexperts' reasoning about assessment. They form the basis of what one might call folk test theory or intuitive test theory. They ground discussions of test theory in the classroom, in the news, and in educational policy. Intuitive test theory works reasonably well for everyday uses like "Friday's math quiz". It fails when you want to design an adaptive test, measure the change in the proportion of students reading Above Basic from a complex large-sample assessment, or base decisions on performance in an interactive computer simulation of problem-solving. The conclusion speculates on prospects for improving the general public's understanding of test theory.

Key words: Assessment, cognitive psychology, expert-novice studies, intuitive physics.

> In most domains of knowledge, we develop very powerful theories when we are very young. School and the disciplines are supposed to reformulate those theories and to make them more comprehensive and more accurate. As long as we stay in school, we can maintain the illusion that the effort has succeeded, but ... once we leave school, the illusion disappears and there is a 5-year-old mind dying to get out and express itself...
>
> No one has to tell a kid that heavy objects fall more quickly than light objects. It's totally intuitive. It happens to be wrong. Galileo showed that it was wrong. Newton explained why it was wrong. But, like others with a robust 5-year-old mind, I still believe heavier objects fall more quickly than lighter objects....
>
> The only people on whom these engravings change are experts. Experts are people who actually think about the world in more sophisticated and different kinds of ways. ... In your area of expertise, you don't think about what you do as you would when you were five years of age. But I venture to say that if I get to questioning you about something that you are not an expert in, the answers you give will be the answers you would have given before you had gone to school.
>
> Howard Gardner, 1993, p. 5

Introduction

The first thing you must know about how humans understand the world is that we make up stories--narratives, as the psychologist Jerome Bruner (1990) calls them. Stories about why people, including ourselves, do what they do, in terms of beliefs, motives, and plans. Stories about how our cars and our computers work, or don't work, in terms of causes, patterns, and linkages.

The second thing you must know is that we make up these stories whether or not we truly understand what is going on. Five year olds, even two years, are driven in exactly the same way as adults and experts to understand in terms of narratives what is happening in the world around them. As Howard Gardner pointed out, their stories can differ, often substantially. Richard Feymann's story for a thrown rock might follow the plot line of "the path of least action" and admit to a rigorous rendering in differential calculus, whereas little Jimmie's story is that the rock wanted to get back down to the ground where it belongs.

The point is that people make up a plausible story, based on what they've experienced themselves and what they've picked up, however loosely or formally, from the culture around them. For example, engineers have acquired concepts such as Newton's laws to reason about physical phenomena such as bridges and rockets, while most five year olds and most senators haven't. Senators are experts in other domains, of course--law, persuasion, how to get things done in Washington--but when it comes to reasoning about physical phenomena, the concepts they have to work with simply aren't as general, as powerful, or as scientific as those of engineers.

The same is true in just about every discipline you can thing of. It is true, I will argue, in educational assessment. To get started, I will review some results from psychologists' investigations into how people who are not experts in physics think about physical phenomena--intuitive physics, as it is called. At its heart are a set of basic beliefs about how the world works, story elements or subplots as it were, called "p-prims" for reasons that will soon become clear. They are definitely not Newton's basics, much less Dirac's or Einstein's, even though they may share some of the words that appear in an expert's compendium. What's surprising is how well they work for the experiences in our everyday lives. They only get us into trouble when we contemplate situations that lay outside their range, in which case expert models are indispensable.

Then we'll apply this kind of thinking to tests and assessments. I'll list a number of p-prims about testing that accord with our everyday experience, and serve reasonably well for many common testing situations; that is, some elements of an intuitive test theory. I will summarize an expert view of test theory--a set of ideas and tools just as powerful and just as strange in their own way as an expert's view of physics, and return to comment on the intuitive test theory p-prims from this perspective. I will conclude with some thoughts on the effects of intuitive test theory on assessment policy and assessment practice, and on prospects of improving the situation.

Intuitive Physics

One consequence of the "cognitive revolution" in psychology that began in the 1960's was a closer look at how people develop expertise in real life activities as varied as chess, radiology, writing, and volleyball. A significant finding across domains is that experts don't just know more facts than novices--although they usually do--but that they organize what they know around deeper principles and relationships. Novices have more fragmented knowledge, related to particular situations or organized around surface features of problems. For example, Paul Feltovich, Micki Chi, and Bob Glaser (1981) asked expert physicists and novices to sort a number of problems into groups. The novices produced piles of spring problems, pulley problems, and inclined plane problems. The experts produced piles associated with equilibrium, Newton's Third Law, and conservation of energy--each containing some spring, pulley, and inclined plane situations.

In 1983, psychologist Andrea diSessa (1983) introduced the notion of 'phenomenological primitives', p-prims for short, to explain non-experts' reasoning about physics. These are primitive notions in the sense that they "stand without significant explanatory substructure or explanation" (diSessa, 1983, p. 15). Familiar examples are "Heavy objects fall faster than light objects", "things bounce because they are 'springy'", and "Continuing force is needed for continuing motion."

Physical p-prims are based on our everyday experience. When we push on a box, it moves; when we stop pushing, it stops moving. Cannon balls really do fall faster than feathers. Physicists know this, of course, but they can drop to a deeper level of explanation when they need to, to the more sophisticated primitives of scientific physics. The distinguishing feature of intuitive physics, or intuitive reasoning in any field (and remember that we *all* reason like this in almost every domain and every activity in which we don't happen to be experts!) is that the p-prims are the bottom line--it's a matter of how far we can go before we have to say "well, that's just the way it is."

Some of the p-prims of intuitive physics borrow words such as force, energy, and momentum, a heritage from the culture and maybe a physics class taken long ago, but the terms are not used in the sense as experts use them. They don't sort the concepts out in the same ways, or embed them in a web of qualitative and quantitative relationships. A set of p-prims is not a coherent system, and a person's set of p-prims can contain some that contradict others. They are employed to reason about physical situations, and a model of sorts is assembled to reason about that situation. The features of a situation tend to call up some p-prims but not others, so a person's intuitive models can be quite different for two situations that are formally equivalent from the point of view of, say, Newton's laws.

The surprising thing is how well intuitive test theory works for guiding everyday action. You can think you are imparting a substance called impetus to the tennis ball when you throw it for your dog, and the ball flies until the impetus wears off. You gauge how much of this substance you want to impart to the ball, and gauge your throw accordingly--and, by golly, the ball goes where you want it to. Your impetus theory is wrong, but neither you, the dog, nor the ball knows this, and the job gets done just fine.

Intuitive physics works good enough for playing catch with your dog or for building a birdhouse. But it doesn't work for building a bridge or shooting a rocket to the moon. One part of becoming an expert in physics is learning these more sophisticated ways of thinking, and another part is knowing when you need to use them. Concepts and relationships outside everyday experience, nonintuitive or even counterintuitive, can then be brought to bear on familiar and unfamiliar situations alike. It is the ones that lie outside everyday experience for which they are ignored at peril.

Some Test-Theory P-Prims

Virtually all of us have taken tests, and many of us have made tests for others to take. For Americans who go to school or hold jobs in the 21st Century, tests are nearly as familiar an experience as pushing boxes and watching things fall. We need to reason about tests, we need to tell stories about them--their purposes and their construction, our performances and our scores--and we need concepts to do so. In a following section I will sketch in a few paragraphs how experts in assessment think about these things. Unless you are an expert in assessment, it is almost certainly not the way you think about them, and some of the ideas will be quite unfamiliar to you. But first I will list a number of beliefs about testing that my colleagues and I come upon time and again in discussions of tests in everyday conversations. We will return to them presently.

- *A test measures what it says at the top of the page.*
- *A test is a test is a test.*
- *Any two tests that measure the same thing can be made interchangeable, with a little equating magic*
- *A score is a score is a score.*
- *You score a test by adding up scores for items.*
- *93% is an A, 85% is a B, 78% is a C, and 70% is passing.*
- *Multiple-choice questions only measure recall.*
- *It's easy to write test items.*
- *You can tell if an item is good by looking at it.*
- *You can tell if a test is good by looking at it.*
- *Technology will solve testing problems by making it possible to get voluminous amounts of data.*

Scientific Test Theory

A scientific perspective on assessment starts by recognizing that assessment isn’t fundamentally about items and scores. These are the springs and pulleys of testing. Rather, assessment is a special kind of evidentiary argument (Messick, 1989, 1994). It is about reasoning from a handful of particular things students say, do, or make, to inferences about what they know, can do, or have accomplished as more broadly construed. So first you need a perspective on the nature of knowledge or skill that’s important, which, for any given student, you can never know with certainty. You need a rationale that connects this view of knowledge, which you can’t see, to things that you can see—maybe right and wrong answers, but maybe problem-solving steps, or justifications for building designs, or comparisons of characters in two novels in terms of transaction theory. You need a rationale for the kinds of tasks or assignments that will elicit this evidence, and an argument for what is meaningful and why in a student’s performance. These arguments that connect tasks and performances in ways that give us clues about what they know and can do--these are the Newton's laws of testing.

A key tool in scientific assessment is using probability-based models to characterize what one knows about students' knowledge and skills from the information in their performances--and just as importantly, what one doesn't know. The use of probability distributions to express our belief about student variables provides a quantitative basis for characterizing the accuracy of measurement, planning test configurations, figuring out how many tasks or raters we need to be sufficiently sure about decisions, and monitoring the quality of large assessment systems. We can extend the probability tools to new kinds of data and assessments, such as ones that adapt tests to individual students in light of how well they are doing or their instructional backgrounds, and tests of problem-solving in computer-based simulations where the problem evolves in response to the student's own actions. These probability models are the calculus of testing, all but unknown to the nonexpert.

It is important to mention that the use of probability models to manage information doesn't restrict the kinds of knowledge and skills we can model. While psychometrics arose around 1900 in the pursuit of measuring traits such as intelligence, the same modeling approach can be applied with all kinds of psychological perspectives and all kinds of data. The variables in student model can be many or few; they can be measures or categories; they can concern knowledge, procedures, strategies, or attunement to social situations; they can be as coarse as "verbal reasoning" or as fine-grained like "being able to map playground situations into the schemas of Newton's laws." What is observed, how it is evaluated, and how it is modeled will depend partly on a psychological perspective and partly on the job at hand. Designing an assessment is like building a bridge. The evidentiary argument and the probability models are like Newton's laws, in that you have to get them right or the structure will collapse. But they aren't sufficient to determine the project. Decisions about location, materials, and design are also driven by the resources you have, the constraints you work under, and the needs of the clients.

The probability models entail another fundamental concept of scientific test theory that is largely unknown outside the discipline: The interplay between the models and the data can tell us when our story is amiss. Charles Spearman's (1904) had the insight a century ago that under the right conditions[1], it is possible to estimate the quantitative features of relationships among both variables that could be observed and others which by their nature never can be--and what's more, to gauge how far and in what ways the data and the posited model disagree. That is, probability-based test theory models are falsifiable, to use the philosopher of science Karl Popper's phrase. Science is just telling stories, but ones that submit to reality checks.

Ms. Pinelas brings little if any of this machinery in constructing, analyzing, and drawing inferences from Friday's math quiz. It doesn't need to be. Familiar testing practices, have evolved into familiar forms of testing that often work good enough for familiar situations if we just follow the usual processes. The principles that make them work in the situations they evolved for are there, invisible, built in the processes and the pieces that we can see. Popular conceptions of how and why familiar assessment work hold the same ontological status as impetus theory--dead wrong in the main, but close enough to guide everyday work in familiar situations. It is when we move beyond the familiar that they break down.

Revisiting the P-Prims

Let's take another look at the list of test theory p-prims presented earlier, this time from the perspective of scientific test theory. We can begin to understand where the p-prims and where they do in fact correspond with our experience, but also see where they will break down.

A test measures what it says at the top of the page.

A score on a purported test of historical analysis can be determined less by how well a student can analyze historical materials than by a host of other factors that influence performance and on which people can differ substantially. Examples are a student's familiarity with what the grader is looking for, whether she happens to be familiar or unfamiliar with a particular topic, familiarity with the kind of test and testing situation, motivation and pressure, knowing the computer interface the test is given on, how well

[1] In statistical terms, if the parameters are identified. Conditional independence is key, because CI relationships enable us to make multiple observations that are assumed to depend on the same unobserved variables in ways we can model.

the test jibes with her previous instruction or coaching, and how well she can plow through the turgid prose the question is written in. The power of our knowledge is its connection to uses, to situations, to other things we know. The value of an assessment--the evidence about a person their performance provides us--depends on how well these factors fit together, and how well we can incorporate it into our inferences.

Intelligence tests are a notorious example. Performance on a particular drop-in-from-the-sky intelligence test does indeed reflect a capability to do some kinds of useful reasoning in some certain circumstances. But there are lots of kinds of intelligent behavior in life, some of which are predicted pretty well by scores on intelligence tests and others of which are not. A person is a good chess player, for example, not because they are smart in a general sense but because through study and practice and playing many games reflectively, they have learned a great deal about the patterns and the strategies in that domain (de Groot, 1965).

A test is a test is a test.

This p-prim is a corollary of the preceding one. Some tests that are called fourth-grade mathematics tests, for example, focus on concepts, others on computations, and still others on using math in real-world situations. They reflect different aspects of what students know and can do with math. Ms. Pinelas can build her quiz so that her students are familiar with the notation, item types, and evaluation standards; a drop-in-from-the-sky can't do this. Assessment projects that require extended work in math can be done in conjunction with instruction over weeks, but they're not well-suited for a single setting drop-in-from-the-sky test. Every assessment test has its own profile of what skills and knowledge it can tell you about, what it costs and what purposes it who provides useful information for, and its implications for students' learning. The same test can be exactly right for one purpose and setting, and a disaster for another. Good assessment designers know this, and design different assessment for different purposes in light of purposes, constraints, and resources.

The dangerous fallacy follows from this p-prim: That you can take a drop-in-from-the-sky test constructed to gauge knowledge in a broad content areas for students about whom you know little else, and by somehow coming up with a different way of scoring it, get diagnostic information that is useful to the classroom teacher for individual small-scale instructional decisions. The problem isn't with the items themselves. Rather it is that you can't match up items and students and classrooms to focus on the question of what to work on next, when that matchup would be different for different classrooms at different points in time.

Any two tests that measure the same thing can be made interchangeable, with a little equating magic.

This is test theory's equivalent of the perpetual motion machine. Why do folks believe it? First, it seems to happen all the time. We are well aware that large scale testing programs like the SAT and the Iowa Test of Basic Skills (ITBS) continually generate new test forms, and psychometricians routinely equate scores on the new forms to scores on the old ones. Secondly, it seems to make sense, as it follows from the preceding p-prims. If you think that tests measure what they say they measure, and that all tests that measure it are pretty much the same, there seems to be no reason why we couldn't make the evidence from different ones somehow equivalent.

But the correspondence between the evidence from one test another superficially similar test is determined by the different aspects of knowledge that different tests tap, the amount of information they provide, and the ways they match better or worse to different students' instruction all impose. The SAT and ITBS can do it from one test form to the next not because of the equating formulas they apply, but because they work so hard to create for every form the a very similar mix of items, to get at the same sets of skills in the same ways with the same difficulties. When tests are not designed up front to be parallel in this way, quantifying how far and in what ways information from one test informs particular questions posed by another, requires expert-level test theory.

With legislation for measuring student progress and establishing common standards for achievement, policy makers are currently interested in linking tests from different states and the National Assessment for Educational Assessment. There is a long and distinguished line of scientific publications pointing out the limitations of linking and equating, including three recent reports by the National Research Council (e.g., NRC, 1999). The idea that the disparate tests can somehow be made equivalent with some equating magic will not go away, however, because life would be much easier if it were true, and under intuitive test theory, there is no reason why it shouldn't be!

A score is a score is a score.

In assessment, as in most inferential problems we encounter, the data we have is rarely perfect and conclusive. And usually we could have obtained different data: more, fewer, or different test items; more, fewer, or different raters; maybe wholly different kinds of data. Once we distinguish what we want to make inferences about--a student's knowledge or skill --from the evidence we have about it, we can gauge how much evidence we have, and compare evidence from real and hypothetical alternatives. This distinction, roughly that of measurement error, is not a natural part of everyday reasoning about testing (although we are more likely to invoke the concept when we do worse on a test than we expected). After all, how could there be a 'truer' score than the score a student actually gets? This p-prim is reinforced by the usual experience that just one test score is obtained, and decisions are made on its basis without 'what-if' considerations concerning hypothetical administrations of alternative measures. This is often good enough for the uses we make of test data, but without the expert model we cannot address the accuracy and validity of those uses. The best way to bring home the existence and the consequences of noise in test scores is to administer more than one, and let people see for themselves the surprisingly large differences that usually result.

You score a test by adding up scores for items.

This is indeed how 99-percent of the classroom quizzes and tests in the world work, and it works just fine for them. One can hardly be blamed for holding this p-prim. But it presumes that the target of inference is students' overall proficiency in some domain of tasks, and the tasks are relatively independent positive indicators of that proficiency. This is the simplest and most familiar case of all the relationships there can be between targets of inference and bits of evidence about it. The problem is that it can't handle the dependencies among more complex forms of evidence and multifaceted models of the knowledge and skill. It fails, for example, for large integrated performances like the National Board of Professional Teaching Standards' videotaped lesson plans and teaching sessions. It fails for interactive problem-solving simulations, in troubleshooting or patient for example, when each action you take takes changes the situation and constrains or facilitates your next action. It fails for collections of tasks that tap different mixes of skills and knowledge, such as language assessments that call for not only vocabulary and grammar, but knowing how to carry out meaningful conversations, use cultural information, and accomplish real-world aims such as bargaining. And it fails for assessments that aim to distinguish conceptions and misconceptions as opposed to correctness, when the target isn't how many problems can a student solve, but how is she thinking so that her responses make sense to her--so we can better figure out what she might work on next to improve her understanding (Figure 1 shows an example from whole number subtraction). These kinds of assessments open the door to more powerful connections between assessment and learning. Scientific test theory is the only hope at designing them effectively and making sense of the data they produce.

93% is an A, 85% is a B, 78% is a C, and 70% is passing.

This p-prim follows from the previous one, with the additional assumption that the tasks that make up a test have been written and selected so that these proportions line up nicely with our degree of satisfaction about how well students have accomplished the goals that were set out for them. It presumes that for all tests, somehow, the same percent correct corresponds in some fundamental way to the same level of performance.

A friend of mine who works on certification and licensing tests told me a state legislature passed a law that the passing score on the plumber's licensing exam--which didn't yet exist--would be 70. As always, my friend and his colleagues worked with experts in the field to determine the kinds of knowledge and skill one needs to be a safe and competent plumber. They worked with plumbers to create a collection of tasks to probe this knowledge. They tried the test out with groups of competent plumbers and apprentices who were not ready to practice on their own. They set a passing score that would best differentiate the two groups. All this work is a sound foundation for creating a valid licensing assessment and setting a justifiable level of performance for a licensing decision. And when they got the number, whatever it turned out to be, they would add or subtract whatever number they had to make the passing score 70.

This p-prim is plausible because so many of the tests we took in our own school careers were reasonable approximations to it. But they didn't get that way by accident. Good teachers who wanted to use this grading scheme thought carefully about what they wanted students to learn, and the conditions under which they could exhibit it. They set up tasks and evaluated them to get data, then looked hard at the data. If the scores they saw from their student didn't jibe with their what they expected, they went back to the drawing board. Items unreasonable or unclear? Revise them or replace them. Students weren't learning what we expected? Improve the lesson, check whether students have the background they need, verify that they were really working.

The problem is that you can make easy and hard tests from the same collection of items, and the same level of knowledge will produce a higher score on the easy test than on the hard test. Item response theory (IRT) psychometric models originated in the 1960's to characterize items in terms of their difficulty and other characteristics, so that students can be administered different sets of items and still be compared on the same scale--harder ones for fifth graders and easier ones for third graders, for example, or, as in the GRE, computer administered tests that are customized to each examinee on the basis of their performance as it unfolds (Wainer et al., 2000). So what is an A now, or a B, or a C? You don't decide on the basis of percentage; you decide on the basis of the performances on the items in the collection that are expected at different points on the scale. This process can be as rigorous as a fixed percentage in situations that call for it. The whole operation can simplify down to what Ms. Pinelas does anyway, when that's appropriate. But the underlying principles provide a deeper understanding of why the standard procedures work in familiar situations, and machinery for creating rather different procedures for new situations--very different arrangements of springs and pulleys, but the same Newton's laws underneath.

Multiple-choice questions only measure recall.

It probably is true that most of the multiple choice questions that people encounter in school only test recall. And it is certainly true for the multiple-choice questions written for someone who believes that's all they can do! But while factual recall items may be the easiest kinds of multiple-choice items to write, it doesn't have to be the case. For example, a multiple-choice test of subtraction items can be written so that patterns of right answers and wrong answers reflect particular buggy procedures, and tell us more about a student's understanding than overall performance on open-ended items would.

Similarly, the research in physics education sparked by work like diSessa's has led to the development of multiple-choice tests that reveal which p-prims students are using. Rather than the usual open-ended computation items, the items on the Force and Motion Conceptual Evaluation (FMCE; Thornton & Sokoloff, 1998) present descriptions of everyday situations and ask students to choose explanations of what is happening or to predict what will happen next. Some alternatives reflect Newton's laws, but others reflect p-prims that are in line with Galileo's thinking, medieval impetus theory, Aristotle's beliefs, or wholly nonscientific justifications. The situations vary in ways that research suggests bring out particular p-prims. Newton's Third Law says that for every action (or force) there is an equal and opposite reaction. If object #1 exerts a force on object #2, then object #2 also exerts an equal and opposite force on object #1. When a car and a small truck of the same weight moving at the same speed collide head-on, most students chose the response that says "The truck exerts the same amount of force on the car as the car exerts on the truck." Okay so far, but this is a canonical example for the third law--easy to give the answer Newton would, without understanding the underlying principle. When the small truck is replaced with a huge semi travelling only half as fast, more students choose "The truck exerts a larger force on the car" because it is larger, or "The car exerts a larger force on the truck" because it is going faster. These responses reflect alternative, in this case conflicting, p-prims.

In and of itself, the format of a task, be it multiple-choice, open-ended, simulation-based, or hands-on performance, doesn't determine the kind of thinking it will elicit in a student. What's more, the same task can give rise to different kinds of thinking in different students, depending on how it fits in with their background and experiences. To a high school algebra student, figuring out the sum of the numbers from 1 to 101 is a simple application of a familiar formula. Rather different cognition was at play when Carl Friedrich Gauss perceived the relationship as an original insight as a seven-year-old. Multiple-choice items can indeed be used to test recall of facts, and that's all most of them do. But when one considers the kinds of concepts or relationships one wants to tap, and the kinds of discriminations that an understanding of them reveals, it is possible to write multiple-choice items that go far beyond recall. The principles for creating such items aren't obvious, though, and aren't a common part of most people's theory of tests.

It's easy to write test items.

Auto insurance surveys tell us that just about every thinks they are a better-than-average driver. The same holds for writing test items. It must be easy, right, since we do it all the time? But creating good assessment tasks isn't something you do in isolation. You think about what knowledge or skill you need to learn about, for the purpose of the assessment is--for remember, different purposes require different kinds of information. What do you need to see students say, do, or make, to give you clues about what the know, can do, of have accomplished? How to you know it when you see? What kinds of materials or tools or support should students have? What conditions and performances best provide evidence, and how might constraints as to time, money, or learning opportunities trade off in assembling tasks? Ironically, the more you know about writing test items, the more challenging it is to write good ones.

You can tell if an item is good by looking at it.

As seen in the discussion of the previous p-prim, knowing if a test item is good depends on things that you can't see when you look at the item on the paper of the computer screen. To be a good test item for a given purpose, there must be a felicitous match among not only what the item provides and requires, but that particular purpose, what the student knows about the context of the item and the scoring rules, and what you know about what the student knows. A bad mismatch at any of these points and the item will fail to provide the evidence you need to do the job, no matter how good it looks on paper. For example, consider an open-ended item for my Advanced Placement calculus class that uses my notation, will be scored with the rubric my students have become used to, and calls for applying what they've been studying for a month now to a real-world situation a lot like another one we discussed in class. A perfect item for me about what they understand of what I want them to learn. A terrible item to include in the Grade 12 National Assessment of Educational Progress, which drops in from the sky and presents tasks to a random sample of students across the country--most of whom are not familiar with my notation and my grading style, and for whom my task wastes ten minutes of valuable testing time.

The converse of this p-prim is pretty much true, though. Some items are so convoluted, substantively wrong in their content, or hard to know how they'll be evaluated that it is difficult to conceive any purpose they will serve. In these cases you *can* tell an item is *bad* just by looking at it.

You can tell if a test is good by looking at it.

As with the previous p-prim, this one again misses the distinction between a test in isolation and a test being used for a particular purpose in a particular context. It also misses the evidentiary-reasoning axiom that the inference you can draw from data depends critically on what else you know and can assume. A teacher can give a highly focused and detailed quiz on the unit she just taught, knowing that this is what the students studied, these are the knowledge representations and evaluation rules they are familiar with, and what they should be familiar with and what is a stretch. Valuable diagnostic information for this teacher will result. The tight focus that makes the test so valuable for this teacher makes it worth less for a drop-in-from-the-sky test. The background knowledge cannot be assumed; there are multiple explanations for poor performance that can't be sorted out because we don't know the relationships between the students and tasks.

As above, you can sometimes tell that a test is so flawed that will be bad for almost any purpose, but you can't tell whether a test will provide good evidence for a given purpose without knowing that purpose.

Technology will solve testing problems by making it possible to get voluminous amounts of data.

Every keystroke, every second, from every student? Sounds great, right? But the challenge is not how much data one gets, but how much sense one can make out the data one gets. The evidentiary reasoning principles noted above are key here, because they are where you think about what kinds of evidence you need for what kinds of inferences, and the connection between them. Gathering huge volumes of data and hoping somebody can figure out what it means later is the hard way to go about doing this, and rarely works well as a practical strategy (although it can be quite appropriate on a small scale in early research, to help understand just what it is you want to do). The reason that multiple-choice tests and add-em-up scoring is still so widely used, despite advances in technology and psychology, isn't because the information multiple-choice tests provide is so great (although it can be better than most people think); it is because people know what to do with this kind of data.

- *Analogy #1: Intelligence analysis.* The National Security Agency can get millions of satellite photos from practically everywhere on earth. Their first problem is how to determine if there is anything important happening with way too few resources and people to study them all. Second problem is that some important things, like peoples' intentions and knowledge, don't show up in satellite photos anyway.
- *Analogy #2: X.* When Spike Lee made a biography of Malcolm X, he didn't put out a 39-year long video of every moment of Malcolm's life; he put together a depiction of incidents in a way that told a story: parts of a life, chosen and structured, to tell a story. A different filmmaker would have made a different film, with different choices to suit his purposes. Assessments are more like this, gathering data, under constraints, within a way of thinking, to serve a purpose.

My prediction is that the big payoffs in assessment from developments in technology, and from developments in learning psychology as well, won't be in doing better jobs at drop-in-from-the-sky large scale testing. They will help some, of course. But I expect the big payoff will be assessment connected intimately with learning, so that focused and purposive use of technology will provide information about the character of a student's knowledge and guidance for what to do next. That information won't just be to policy-makers in big, slow, decontextualized feedback loops, but smaller and faster feedback loops with more direct bearing on learning, to the teacher, to the students themselves, or to adaptations for the learning environment directly.

Discussion

Intuitive physics works okay for building a birdhouse or playing catch with your dog, but not for building a bridge or shooting a rocket to the moon. If you want to shoot a rocket to the moon, sooner or later somebody is going to have to do some calculus. In the same way, intuitive test theory works okay for classroom testing and the quizzes in *Seventeen* magazine, but it gets you into trouble when you want to do measure value added in classrooms, evaluate performances on simulation-based tasks, run secure and accurate high stakes testing programs, or measure change in populations using an achievement survey like the National Assessment for Educational Progress (NAEP). It is a serious problem that assessment policy is made on the basis of intuitive test theory.

There is another similarity between intuitive test theory and intuitive physics that has implications for assessment policy. As one's understanding of physics goes deeper into the domain, the concepts and tools depart from everyday physics. The same is true with assessment design and analysis at the frontiers, such as NAEP and simulation-based assessments. People generally accept that this is the case in physics. Experts are consulted, and their perspective becomes part of the policy debate. They don't make the decisions, and they shouldn't; in any social setting, there are more considerations than purely technical ones. But policy discussions should be at least restricted to options that accord with Newton's laws of motion. This is not the case in assessment. P-prims are strongly held and widely held, and there is little compulsion to change.

What can those of us in the technical end of assessment do about this situation? Two aspects of the job that bear on this problem. One isn't fun, but the other is. The one that *isn't* fun is trying to critique or implement policies and programs that have been put together on the basis of intuitive test theory. This kind of project requires a lot of telling people that what they want to do won't work, and to do it right is harder or takes longer or isn't as accurate as they want. The one that *is* fun is working on projects from first principles, especially when you push the frontiers in ways that use new ideas from technology or psychology. You use the machinery of evidentiary reasoning, you design the tasks and the analysis to be in concert, and you accomplish things that couldn't have been done otherwise--and certainly couldn’t have done with intuitive test theory. These existence proofs are the most compelling argument for test theory as a scientific discipline, and it is by spinning off these examples that practice will advance.

References

Bruner, J. (1990). *Acts of meaning.* Cambridge: Harvard University Press.

Chi, M.T.H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

de Groot, A.D. (1965). *Thought and choice in chess.* The Hague: Mouton.

diSessa, A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A.L. Stevens (Eds.), *Mental models* (pp. 15-33). Hillsdale, NJ: Erlbaum.

Gardner, H. (1993). *Educating the unschooled mind.* Washington, D.C.: Federation of Behavioral, Psychological, and Cognitive Sciences.

National Research Council (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Committee on Equivalency and Linkage of Educational Tests, M. J. Feuer, P. W. Holland, B. F. Green, M. W. Bertenthal, & F. C Hemphill (Eds.). Washington DC: National Academy Press.

Thornton, R.K., & Sokoloff, D.R. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation. *American Journal of Physics, 66,* 228-351.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (second edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

$$\begin{array}{r} 821 \\ \underline{-\ 285} \\ 664 \end{array} \qquad \begin{array}{r} 885 \\ \underline{-\ 221} \\ 664 \end{array}$$

$$\begin{array}{r} 63 \\ \underline{-\ 15} \\ 52 \end{array} \qquad \begin{array}{r} 17 \\ \underline{-\ 9} \\ 12 \end{array}$$

Figure 1: Responses consistent with the "subtract smaller from larger" bug. When the 'subtract smaller from larger' bug is present in a student's configuration of production rules, problems requiring borrowing will show the characteristic pattern of incorrect responses that results from simply subtracting whichever number in a column is smaller from whichever is larger. When borrowing is not required, this bug does not affect responses; they will be correct or incorrect in whatever ways are consistent with the student's other rules. Knowing how a student is thinking, as opposed to simply how many items they are getting right, provides better information to help them improve.

.