Intuitive Test Theory

CSE Report 631

Henry I. Braun
Educational Testing Service

Robert J. Mislevy
CRESST/University of Maryland, College Park

May, 2004

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.6 Study Group on Cognitive Validity
Strand 1 Cognitively Based Models and Assessment Design
Robert J. Mislevy, Project Director, CRESST/University of Maryland, College Park

The work reported herein was supported by the Educational Testing Service and under 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.

INTUITIVE TEST THEORY[1]

Henry I. Braun

Educational Testing Service

Robert J. Mislevy

CRESST/University of Maryland, College Park

Abstract

Psychologist Andrea diSessa coined the term "phenomenological primitives", or p-prims, to talk about nonexperts' reasoning about physical situations. P-prims are primitive in the sense that they stand without significant explanatory substructure or explanation. Examples are "Heavy objects fall faster than light objects" and "Continuing force is needed for continuing motion." P-prims are based on experience. They are not a coherent system; they may even contradict one another. People assemble from them a model of sorts to reason about a given situation. Intuitive physics is wrong from a physicist's point of view, but it works just fine play fetch with your dog or push a couch across the room. It fails when you want to build a skyscraper or send a rocket to the moon. This paper considers p-prims that underlie reasoning about assessment, the basis of what one might call intuitive test theory. Examples are "A test measures what it says at the top of the page," and "Scores from any two tests can be made interchangeable, with a little equating magic." Testing p-prims underlie discussions of test theory in the classroom, in the news, and in policy-making. Again, intuitive test theory works reasonably well for everyday uses like Friday's math quiz. It fails when you want to design an adaptive test, or measure the change in the proportion of students reading Above Basic from a matrix-sampled assessment such as NAEP.

Key words: Assessment policy, p-prim, test theory.

[1] The author order is alphabetical; both authors contributed equally. The second author presented a version 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 this opportunity, and to Neal Dorans, Paul Holland, and Howard Wainer for stimulating conversations on the topic.

Intuitive Test Theory

> 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

Along with making sure that our bodily needs are met, one of our first tasks upon entering the world is to try to make sense of it. We do so by continuous observation and generalization, as well as by absorbing the norms of the culture in which we find ourselves. Our understandings typically take the form of stories—narratives, as the psychologist Jerome Bruner (1990) calls them. These are stories about why people do what they do, in terms of beliefs, motives, and plans.

This mode of developing and retaining understanding carries over to the physical world, natural or man-made. We hear thunder and see lightning, observe objects being thrown and falling to the ground, cars and computers working (or not) and construct stories in terms of causes, patterns, and linkages. Now, we make up these stories whether or not we truly understand what is going on. Adults are driven, in exactly the same way as five-year-olds are, to express their understanding of what is happening around them in terms of narratives.

As Howard Gardner (1993) points out, stories can differ, often substantially. Richard Feynman's story for a thrown rock might be based on the principle 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 construct plausible stories for actions and events, based on what they've experienced themselves and what they've picked up, however loosely or informally, from the culture around them. The Gardner quote highlights two other aspects of these narratives. The first is their tendency to persist, even in the face of evidence to the contrary or confrontation with methods of analysis that are much more powerful. Bruner (1990) makes the same point with respect to what he calls "folk psychology." He defines folk psychology as a system by which people organize their experience in, knowledge about, and transactions with the social world. We learn our culture's folk psychology along with language and social behavior. He asserts that "Folk psychology changes but is not displaced by scientific psychology." (p.14). It is the persistence of these narratives (say, in physics) that can be so frustrating for teachers!

The second is that expertise is often very narrowly focused; that is, outside one's area of specialized training, it is uncommon to do better than a five-year-old. Indeed, the situation may be even more dire. In a now classic study, the psychologists Amos Tversky and Daniel Kahneman (1971) questioned a large number of research psychologists on various aspects of probability and statistics (the design of experiments and the interpretation of the results) that would ordinarily be relevant to their work. Surprisingly, a majority of the respondents harbored naïve (and incorrect) beliefs that, presumably, influenced how they conducted their research.

What is true of psychology or physics is also the case in just about every discipline you can think of. It is true, we will argue, in educational assessment. To set the stage, we will first review some results from investigations into how people who are not experts in physics think about physical phenomena. At the heart of what is often termed "intuitive physics" are a set of basic premises about how the world works -- story elements or subplots as it were, called phenomenological primitives (or p-prims, for short; diSessa, 1983). They are definitely not Newton's basic physics, much less Dirac's or Einstein's, even though they may share some of the words that appear in an expert's compendium.

Perhaps it is not surprising that p-prims and the narratives in which they are embedded work well enough for most situations in our everyday lives. After all, they are grounded in the experiences of many people over many, many years. They can lead to trouble, though, when employed in situations that lie outside their range, in which case expert models are indispensable. Unfortunately, unlike prescription drugs, p-prims (in physics or other disciplines) are usually not accompanied by warning labels with contraindications for use. In a fast changing world, it is increasingly likely that we will find ourselves relying on p-prims that are not up to the job.

To fix ideas, the following section offers a brief tour of intuitive physics, a domain rich in research into the contrasts between everyday reasoning and expert reasoning. We then move to tests and assessments, beginning with a list of p-prims about testing that accord with our everyday experience and serve reasonably well for many common testing situations. These p-prims constitute the elements of an "intuitive test theory." Next, we present a quick summary of an expert's 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. From the perspective of expert test theory, we then comment on the p-prims of intuitive test theory, and provide some examples of testing p-prims in action. In the concluding section, we offer some thoughts on the effects of intuitive test theory on assessment practice and policy, as well as on the 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 radiology, writing, chess, and volleyball. A significant finding across domains is that experts don't simply 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, Micki Chi, Paul Feltovich, 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. The experts' categorization leads directly to solution strategies for the problems.

In 1983, psychologist Andrea diSessa (1983) introduced the term 'phenomenological primitives', or p-prims, 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. A box moves when we push it, and it stops moving when we stop pushing. Cannon balls really do fall faster than feathers. Physicists know this, of course, but, when necessary, they can appeal to a deeper level of explanation, to the more sophisticated primitives of scientific physics. The distinguishing feature of intuitive physics (or intuitive reasoning in any field, remembering 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 go before we have to say "well, that's just the way it is."

Some of the p-prims of intuitive physics use words such as force, energy, and momentum, a legacy from the culture or a physics class taken long ago, but the terms are not employed in the same way that experts use them. They don't sort concepts in the same ways, or embed them in the same 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 easily contain some p-prims that contradict others. They are employed to reason about physical situations, and a model of sorts is assembled to address a given situation. The surface features of a situation tend to elicit 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 p-prims work 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 estimate 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 well enough for playing catch with your dog or for building a birdhouse. But it doesn't work for constructing a bridge or shooting a rocket to the moon. One aspect of becoming an expert in physics is learning more sophisticated ways of thinking, but another is knowing when you need to use them, and yet another is recognizing when they fail. (Science is also about telling stories, but ones that submit to reality checks.) In scientific physics, concepts and relationships that may be nonintuitive, or even counterintuitive, can be brought to bear on familiar and unfamiliar situations alike. Individuals facing challenges that lie outside everyday experience ignore scientific physics at their peril!

Some Test-Theory P-Prims

To Americans who go to school or hold jobs in the 21st century, taking tests is nearly as familiar an experience as pushing boxes and watching things fall. We need to tell stories about them—their purposes, their construction, our performances on them—and so we need concepts to do so. In the next section, we will briefly sketch how experts in assessment think about these aspects of tests. Unless you are an expert in assessment, it is probably not the way you think about them, and some of the ideas may be quite foreign to you. But first we list a number of beliefs about testing that we, and our colleagues in educational measurement, encounter time and again in everyday discussions involving tests. 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.
- A score is a score is a score.
- Any two tests that measure the same thing can be made interchangeable, with a little "equating magic"
- 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.

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

- Multiple-choice tests = standardized tests = high-stakes tests

Scientific Test Theory

A scientific approach to assessment recognizes that, fundamentally, assessment isn't about items and scores. These are like the springs and pulleys of testing. At a deeper level, assessment is a special kind of evidentiary argument (Kane, 1992; Messick, 1989, 1994; Mislevy, in press): Assessment is about reasoning from a handful of particular things students say, do, or make, to more broadly-cast inferences about what they know, have accomplished, or are apt to do in the future.

The starting point for an application of scientific test theory is a clear understanding of the purpose of the assessment and a perspective on the nature of the knowledge or skills that are the focus of attention. Next is the link between this view of knowledge and skills, which you can't see, to things that you can see—right and wrong answers, problem-solving steps, justifications for building designs, or comparisons of characters in two novels in terms of transaction theory, to cite just a few examples. This analysis resolves into making a case for what is meaningful – and why—in a student's performance, in light of the purpose of the assessment. A rationale is also required for the kinds of assignments or challenges that will elicit the evidence to support the intended inferences about students. Conceptual links connect tasks to student performances to judgments about what they know and can do. These are the testing counterparts of Newton's laws in testing.

Now, Newton's laws of motion are deterministic; that is, given a complete description of an object (e.g., its mass, current position, and velocity) we can calculate exactly the effect on its motion of an application of a particular force. In test theory, we can formulate a student model that describes one or more aspects of a student's knowledge or skills. Since the components of the student model cannot be observed directly, we have to use probability theory to express our beliefs about the likely values of these components. As we accumulate more data about the student, we can employ the calculus of probabilities to update our beliefs.

The use of probability-based models to describe what we know, and what we don't know, about a student is a key tool in scientific assessment. It provides a quantitative basis for planning test configurations, calculating the accuracy and reliability of the measurement process, figuring out how many tasks or raters we need to be sufficiently sure about the appropriateness of decisions based on test scores, or monitoring the quality of large-scale assessment systems. We can also apply the tools of probability to new kinds of testing processes, such as ones that select discrete tasks to present to individual students in light of how well they are doing or their instructional backgrounds, or tests of problem-solving in computer-based simulations where the problem itself evolves in response to the student's actions. These probability models and their essential role in reasoning are all but unknown to the nonexpert.

It is worth pointing out 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 with the goal 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 the 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, and how it is modeled and evaluated, will depend partly on a psychological perspective and partly on the job at hand. Designing an assessment is like building a bridge. The evidentiary arguments 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. In architecture and engineering, decisions about location, materials, and various features of the design are strongly influenced by the resources available, the situational constraints, and the needs of the clients. Similar processes are at work in measurement (Braun, 2000).

The use of probability models is responsible for a critical aspect of scientific test theory that is largely unknown outside the discipline; namely, that the interplay between the models and the data can tell us when our story is amiss. Charles Spearman (1904) had the insight that, under the right conditions[2], it is possible to estimate the quantitative features of relationships among both variables that could

[2] In statistical terms, if the parameters are identified. Conditional independence (CI) 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.

be observed and others which, by their nature, never can be. What's more, it is possible to gauge how far and in what ways the data and the posited model disagree. That is, the models of probability-based test theory are falsifiable, to use the philosopher of science Karl Popper's phrase. Test theory is indeed a branch of science.

The typical classroom teacher brings little if any of this machinery in constructing, analyzing, and drawing inferences from Friday's math quiz. Usually this is perfectly fine, appropriate to the purpose and the context. Assessment practices have evolved into familiar forms of testing that often work well enough in common situations. The principles that account for why they work in the situations for which they evolved are there—invisible but built into the processes and the pieces that we can see. Popular conceptions of how and why familiar tests work hold the same ontological status as impetus theory—dead wrong in the main, but close enough to guide everyday work in familiar settings, as long as we have the right intuitions about what is important for students to be able to do and how we know it when we see it. It is when we move beyond the familiar that testing p-prims can betray us.

P-Prims Under Scrutiny

We now consider the test theory p-prims listed earlier. Using the insights of scientific test theory, we examine the p-prims, to begin to understand how they might have arisen and where they can break down. In what follows, we sometimes use the phrase “drop-in-from-the-sky” to describe a test—by which we mean a test that is developed outside the school context. The term is meant to connote the remoteness of the test from the day-to-day experiences of the students.

A Test Measures What it Says at the Top of the Page.

It is natural to assume that a name carries meaning. Thus, we expect a test that is called a history test, for example, to measure a student’s accomplishments or proficiency in history. However, a student’s score on such a test can be determined less by how well a student can analyze or interpret historical materials than by a host of other factors that also influence performance, and on which individuals can differ substantially. Examples are a student's familiarity with the testing situation, the kind of test and mode of administration, and even what the grader is looking for.

A common manifestation of this p-prim is making inferences from test scores that extend well beyond what can be reasonably supported. Perhaps the most notorious example is the over-interpretation of the results from standardized intelligence tests. Performance on a particular drop-in-from-the-sky intelligence test does typically indicate a capability to do productive reasoning in certain circumstances. But there are many kinds of intelligent behavior, some of which are predicted pretty well by scores on intelligence tests and others which are not (Gardner, 1983; Sternberg, 1988). People are good chess players, for example, not because they are intelligent in a general sense, but because through study, practice, and reflecting on their performance in many, many games, they have learned a great deal about the patterns and successful strategies in the domain of chess (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 more 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. Furthermore, a classroom teacher can build her quiz assuming that students are familiar with her notation, item types, and evaluation standards. This is more difficult for a drop-in-from-the-sky test. Moreover, assessments that are projects requiring extended work in math can be done over a period of time as part of a program of instruction, but they aren't well suited for a drop-in-from-the-sky test that occurs on a single day.

Each assessment can be described in terms of the skills and knowledge it can tell you about, how much information it provides, its implications for learning, how closely it corresponds to students' background and instruction, and its demands on resources such as equipment, cost, and student and teacher time. The trick is to match a test, with all its many characteristics, with the purpose of testing and the context in which it will be used. Getting the proper match can be a delicate balancing act. For any number of reasons, the same test can be exactly right for one purpose and situation, but quite useless for another. Good test developers know this, and design different assessments for different purposes in light of the characteristics of the students, the available resources, and the constraints of the setting.

A particularly dangerous fallacy follows from this p-prim: That you can take a drop-in-from-the-sky test constructed to gauge knowledge in broad content areas for students about whom you know little else and, by coming up with a different way of scoring it, obtain diagnostic information that will be useful to the classroom teacher for individual small-scale instructional decisions. This generally doesn’t work, and the problem isn't with the items or the scoring rules. It is that effective information about what to do next requires assessment that takes into account what a teacher already knows about a student and provides information in terms of instructional options—not necessarily better items or more items, just the right items for the right student at the right time. Good diagnostic information results from good match-ups, not from good “one-size-fits-all” tests.

A Score is a Score is a Score

With all the criticism that testing attracts, it is remarkable how much credence is typically attached to a single test score. After all, the reasoning goes, how could there be a 'truer' score than the score a student actually gets? This p-prim is reinforced by the usual experience that decisions are made on the basis of a single test score, without 'what-if' considerations concerning the scores that could have been obtained in hypothetical administrations of alternative measures. Indeed, measurement experts recognize that different data could have arisen: by testing on other occasions; with more, fewer, or different test items; employing more, fewer, or different raters (Cronbach, et al., 1972); the variation among what would have been equally good measures of the targeted knowledge or skill ought to taken into account in the use of any of those scores. Perhaps the best way to bring home the concept of “noise” in test scores is to administer multiple tests and let people see for themselves the surprisingly large differences that result.

Once we decide what we want to make inferences about from the data available, using scientific test theory we can gauge how much evidence we have, and compare it with what might have occurred under a variety of hypothetical alternatives. This concept, roughly that of measurement error, is not a natural part of everyday reasoning about test scores (with the major exception that occurs when someone’s score is lower than he or she expected!) Assessment data is not perfect. Reliance on a single score without regard to the uncertainty attached to it may be good enough for typical, low-stakes applications but can be problematic for more consequential ones. Without scientific test theory, we can neither quantify that uncertainty nor evaluate the validity of the use of the test score in a particular setting.

Any Two Tests that Measure the Same Thing Can Be Made Interchangeable, with a Little "Equating Magic"

This is intuitive test theory's equivalent of the perpetual motion machine. Why do people believe it? First, it seems to happen all the time. Almost everyone knows that large scale testing programs like the SAT I 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, because it follows from the preceding p-prims. If you think that tests measure what they say they measure, that all tests that measure it are essentially the same, and you don't concern yourself with measurement error, then there is no apparent reason not to treat evidence from different tests as more or less equivalent.

But the strength of the correspondence between the evidence from one test to another, superficially similar test is determined by the different aspects of knowledge and skills that the two tests tap, the amount and quality of the information they provide, and how well they each match to the students' instructional experiences. The SAT I and ITBS testing programs can do this, not so much because of the equating procedures they carry out, but because they expend considerable effort to create test forms with very similar combinations of questions (item types, content areas, mix of difficulties), in order to tap the same sets of skills, in the same ways. When tests are not designed to be "parallel" in this way, quantifying in what ways information from one test can be used as if it came from another, requires expert-level (scientific) test theory. Some inferences across tests will work well and others will fail. Only with mathematical models can we establish how to carry out linkages, study the nature of biases and magnitudes of inaccuracies for different questions we might ask of the data, and determine which inferences across tests are appropriate in light of the consequences of the decisions to be made.

With legislation mandating the measurement of student progress and the establishment of common standards for achievement, policy makers have expressed considerable interest in linking tests from different states or different test publishers to each other, and to the National Assessment for Educational Progress (NAEP). There is a long and definitive line of scientific publications pointing out the very real limitations to linking and equating different tests with the same name (e.g., National Research Council, 1999). Unfortunately, the notion that disparate tests can somehow be made equivalent with the application of equating magic will not die because life would be much easier if this were true—and under intuitive test theory, there is no reason why it can’t be done!

You Score a Test by Adding up Scores for Items

Almost all classroom quizzes and tests are graded in this way—and it works just fine for them. Consequently, one can hardly be blamed for holding this p-prim. But it presumes that the target of inference is a student's overall proficiency in some domain and that the tasks on the test are relatively independent positive indicators of that proficiency. Indeed, this is the simplest (and most familiar) case of a relationship between targets of inference and bits of evidence about it. However, when interest focuses on dependencies among more complex forms of evidence and multifaceted models of knowledge and skill, the “natural” approach to scoring is severely deficient.

This approach fails for large integrated performances like the National Board of Professional Teaching Standards' videotaped lesson plans and teaching sessions, because multiple, interconnected judgments across many parts of the work are required. It fails for interactive problem-solving simulations (e.g., troubleshooting or patient management), because each action taken changes the situation, and constrains or facilitates the next action. It fails for collections of tasks that tap a variety of skills and knowledge in different mixes, such as language tests that assess not only vocabulary and grammar, but also how to conduct meaningful conversations, use cultural information, and accomplish real-world aims such as bargaining. Patterns of what is done well and where performance is inadequate are required, with the added complication that people trade off their strengths against their weaknesses when they use language in real life. And it fails for assessments that aim to distinguish conceptions and misconceptions (as opposed to correctness); that is, when the goal isn’t to count how many problems a student can solve but, rather, to develop a useful description of her thinking—so that we can better decide what she might work on next to improve her understanding. Figure 1 shows an example from whole number subtraction.

In all of these cases, simple scoring rules don’t make the “grade” because they extract only a part of the evidence contained in students’ responses—sometimes completely missing the patterns that are most important—and therefore can’t support the nuanced inferences that are desired. Scientific test theory, extended and elaborated as needed to deal with new kinds of data and new kinds of inferences about students, is the best foundation for both effectively designing these more complex assessments and making sense of the data they produce.

$$\begin{array}{r} 821 \\ \underline{-\ 285} \\ \mathbf{664} \end{array} \qquad \begin{array}{r} 885 \\ \underline{-\ 221} \\ \mathbf{664} \end{array} \qquad \begin{array}{r} 63 \\ \underline{-\ 15} \\ \mathbf{52} \end{array} \qquad \begin{array}{r} 17 \\ \underline{-\ 9} \\ \mathbf{12} \end{array}$$

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 (VanLehn, 1990). 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.

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 so that these percentages line up nicely with the traditional percent-correct metric of satisfaction for how well students have done on tests of materials that were specifically matched to their instruction. It presumes that somehow, for all tests and all uses and all students, the same percent correct corresponds to the same level of performance.

A colleague who works on certification and licensing tests relates that a state legislature passed a law mandating that "the passing score on the plumber's licensing exam will be 70". Following good test design practice, our colleague worked with plumbers to determine the kinds of knowledge and skills needed to be a competent plumber, one who is able to ply the craft ably and with due regard to safety. The committee then created a collection of tasks to probe the targeted knowledge and skills, pilot tested them with groups of competent plumbers and with apprentices who were judged to be not ready to practice on their own. A passing score was selected to best differentiate the two groups. This is a sound foundation for creating a valid licensing assessment and setting a defensible level of performance for a high-stakes decision. When they got that number, it shouldn’t have mattered what its numerical value was—within the constraints of the testing program—it was by construction a valid cut point to distinguish who should obtain a license and who should not. As a final step, however, the test developers add or subtract a constant to make the passing score exactly 70!

This p-prim is plausible because for many of the tests we took in school, this grading scheme is not a bad choice. But that didn't happen 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 numbers. If the scores they saw from their students didn't jibe with their expectations, they went back to the drawing board to figure out why. Were the items unreasonable or unclear? Is so, they would then revise or replace them. Were the students just not learning what was intended? If so, they would then check whether the students have the background they need, verify that they are really working, improve the pedagogy, and so on.

The difficulties encountered in applying this p-prim and the previous one in more complex settings have led to advances in measurement theory. Indeed, it is possible to construct both 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 features, 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 now is an A, or a B, or a C? You can't decide by just calculating the percentage of correct answers; you should decide on the basis of the pattern of correct and incorrect answers, taking into account the relative difficulty of the items presented, and examining the kinds of tasks that students along different points of the scale can and cannot accomplish. As always it is a value judgment. Now, however, the judgment is explicit rather than confounded with the particulars of items and measurement scales. Now the underlying principles provide a deeper understanding of why the standard procedures work in familiar situations, as well as the machinery for creating new procedures for novel situations—very different arrangements of springs and pulleys, but the same Newton's laws underneath.

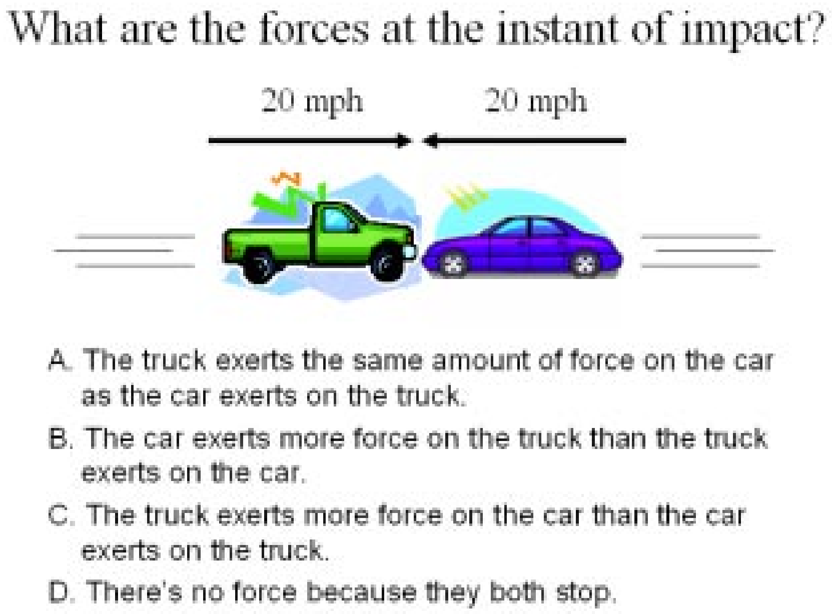
Multiple-Choice Questions Only Measure Recall

This p-prim is often stated as an epithet, part of a comparison to open-ended tasks. Certainly most multiple choice questions that people encounter in school do only test recall—and it is surely true for multiple-choice questions written by someone who believes this p-prim! But while factual recall items may be the easiest kinds of multiple-choice items to write, other types are certainly possible. For example, a multiple-choice test of subtraction can be written so that patterns of right answers and wrong answers, reflecting particular misconceptions or buggy procedures, tell us more about a student's understanding than a total score on a test comprising open-ended items.

Similarly, 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 and modeling items, the items on the Force and Motion Conceptual Evaluation (Thornton & Sokoloff, 1998) and the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992) present descriptions of everyday situations and ask students to choose explanations of what is happening or predict what will happen next. Some alternatives reflect Newton's laws, but others reflect p-prims that are more consistent 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.

For example, 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 (Figure 2), 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 pickup truck is replaced with a huge semi traveling only half as fast (Figure 3), 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, and in this case, conflicting p-prims.

In and of itself, the format of a task, be it multiple-choice, open-ended, or hands-on performance, doesn't fully 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 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 cognitive processes were at play when the seven-year-old Karl Friedrich Gauss perceived the relationship as an original insight. Multiple-choice items can be used to test recall of facts, and indeed most of them do. But if one has clearly in mind the concepts or relationships one wants to probe, as well as the kinds of discriminations that an understanding of them entails, it is possible to write multiple-choice items that go far beyond recall.


What are the forces at the instant of impact?

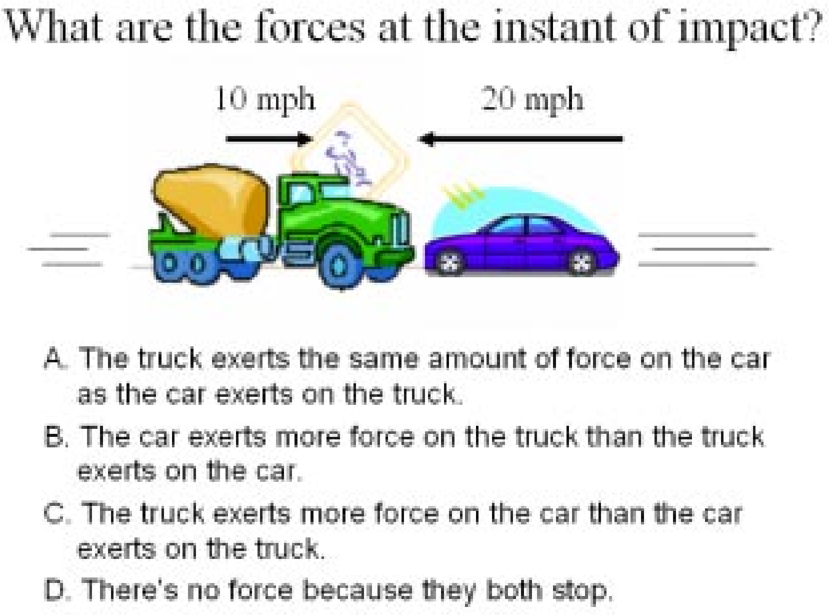
A. The truck exerts the same amount of force on the car as the car exerts on the truck.

B. The car exerts more force on the truck than the truck exerts on the car.

C. The truck exerts more force on the car than the car exerts on the truck.

D. There's no force because they both stop.

Figure 2. A multiple-choice item meant to reveal students' thinking about a canonical Newton's Third Law situation.



What are the forces at the instant of impact?

A. The truck exerts the same amount of force on the car as the car exerts on the truck.

B. The car exerts more force on the truck than the truck exerts on the car.

C. The truck exerts more force on the car than the car exerts on the truck.

D. There's no force because they both stop.

Figure 3. A multiple-choice item meant to reveal students' thinking about a non-canonical Newton's Third Law situation.

You Can Tell if an Item is Good by Looking at it.

This p-prim, as do most of the others, rests on the assumption that items and tests are simple objects whose essence can be grasped by their surface characteristics. However, for an item to serve a given purpose, there has to be a reasonable coherence among that particular purpose, what the item provides and what it requires, the student's understanding of the context of the item and the scoring rules, and what else the assessor knows about what the student knows. A mismatch at any point, and the item may fail to generate the evidence needed, no matter how "good" it looks.

For example, consider an open-ended item devised by a teacher for her Advanced Placement calculus class that uses her notation, will be scored with the rubric her students have become used to, and calls for applying what they've been studying for the last month to a real-world situation that is similar to one discussed in class. This is an ideal probe to elicit their understanding of an important learning objective. However, it would be a poor 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—many of whom would not be familiar with the notation or the grading rubric. Ten minutes of valuable testing time would be wasted for almost everyone confronted with this question. (The converse of this p-prim is more nearly true: You can often tell an item is bad just by looking at it. Logical flaws and confusing instructions, for example, will keep an item from providing useful information for almost any purpose.)

That the appropriateness of an item depends on "more than meets the eye" implies that it is more difficult than most people would imagine to write good items and match them to the contexts in which they will be used. In addition to having a coherent conceptual framework and a strong evidentiary perspective, item writers must also work under constraints of time and money as they build tasks and assemble tests. It is not a vocation for the faint of heart or the novice, as recent missteps in many high-stakes state tests attest. Ironically, the more one knows about writing test items, the more challenging it is to write good ones.

Multiple-Choice Tests = Standardized Tests = High-Stakes Tests

Many of the highly visible tests used today for admissions, licensure, and certification, as well as state accountability for public schools, are alike in three important ways: they have meaningful consequences for students or schools, they are presented under standard conditions, and they use multiple-choice items. This configuration occurs often enough that these three distinct properties are conflated in the public eye so that the adjectives 'multiple-choice', 'standardized', and 'high-stakes' are thought to be synonymous—all ways to describe the same familiar package.

But high-stakes tests can be less standardized and require performances, as is the case with doctoral dissertations and solo flights for pilot certification. Multiple-choice items are found as often in low-stakes classroom quizzes are they are in high-stakes assessments. And standardization is not an all-or-nothing quality. For each aspect of an assessment, there are many choices about how to make the experience similar or different for different examinees—identical, tailored, individual choice? As always, the answers involve tradeoffs. Greater similarity across examinees in some facets tends to support comparisons and facilitate communication of results across time and distance. More individualization allows the tests to be better targeted to individuals' circumstances, although the interpretation of results is more tightly bound to those circumstances. Debates about accountability tests should move beyond this p-prim, as sensible policy demands a more careful distinction among properties of assessments, coupled with models and methods for building tests that optimize their functioning for different purposes and contexts.

Discussion

While intuitive test theory is sufficient for classroom testing and the quizzes in the weekend magazine in the newspaper, it gets you into trouble when you want to evaluate performance on simulation-based activities, run a high-stakes testing program, or measure change in populations using an achievement survey like the National Assessment for Educational Progress (NAEP). There is a strong similarity—and an important accompanying difference—between intuitive physics and intuitive test theory that has implications for assessment use and policy. What's similar is that as understanding and expertise in physics deepen, the concepts and tools depart from everyday physics. The same is true with assessment design and analysis at the frontiers, such as simulation-based assessments and NAEP.

It is generally accepted that this is the case in physics and, moreover, that the complexity must be confronted if one is embarking on a serious undertaking. We have already referred to the paradigmatic example of launching a rocket to the moon. In fact, in 1961,when President Kennedy made his famous promise that by the end of that decade the U.S. would send a man to the moon and return him safely back to Earth, his staff had already consulted with experts about the feasibility of such an endeavor. Two points are noteworthy. First, everyone expected that all the options that would be considered would be in accord with Newton's laws of motion, not Aristotle's. Second, President Kennedy did not assert that, on its flight to the moon, the rocket would have to meet specific milestones that he and his advisors deemed appropriate.

In most issues that involve technical considerations, experts are consulted, and their perspectives become 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 options should be restricted to those that accord with basic principles and broadly held standards of practice—the analogs of Newton's laws of motion.

Unfortunately, this is often not the case in assessment, as a review of the testing policies in many states and the legislative history of the No Child Left Behind Act demonstrate. As assessment-based accountability becomes a more prominent feature of education policy, those standing on the technical side of assessment must confront the reality that critical decisions are made and regulations are drafted on the basis of intuitive test theory, with untoward consequences a likely result. The advent of technology-based assessment may, in many ways, exacerbate the problem. No doubt voluminous data will be produced, but insight will still be in short supply. In fact, a disciplined application of the principles of evidentiary reasoning to design, development, and analysis will be all the more necessary if the investment in technology is to yield meaningful returns.

We remain, then, with the problem that p-prims are both widely held and persistent. We can speculate on the reasons for their persistence. One is that p-prims correspond to a world that is, in many ways, much simpler and easier to comprehend than our own. As a species, we are no fans of complexity and usually must be dragged kicking and screaming toward an admission that the world is really stranger than we would like it to be. Second, p-prims are often embedded in narratives that, by virtue of their structure and their frequent usefulness in everyday experience, resonate deeply and are difficult to dislodge. Third, there are relatively few settings in which testing p-prims (particularly) are directly challenged and experts' conceptions are presented in an understandable manner.

We should note that clinging to narratives unsupported by evidence is not confined to the lay public. In the early part of the 20th century, many psychologists held that parents should not comfort their children for, to do so, would be to encourage an unhealthy dependence that would result in their development into dysfunctional adults. Though empirical support was weak, these views had a substantial effect on child raising. It was only with the justly famous series of experiments carried out by Harry Harlow and his associates on rhesus monkeys that the prevailing dogma was effectively challenged and, eventually, overturned (See Blum, 2003).

What, then, should those of us in educational measurement do? There are at least three lines of attack, one negative and two positive. First, we should not shy away from critiquing policies and programs that are based on intuitive test theory. This involves telling lots of people (some very important) that what they want to do just won't work and that to do it right is harder or takes longer than they might like. Good examples of this sort of activity is provided by Robert Linn in his published work (e.g., Linn, 2000; Linn, 2003), as well as in his tireless efforts to impart basic measurement principles to a variety of audiences. Another example is the report of the NRC Panel addressing the issue of linking state tests to NAEP (NRC, 1999).

A second approach is to use scientific test theory, in conjunction with developments in psychology and technology, to achieve goals that could not have been accomplished otherwise—certainly not by relying on intuitive test theory. These existence proofs are the most compelling argument for test theory as a scientific discipline and for its utility in the setting of education policy. Finally, we need to do a much better job of communicating to a variety of audiences the basics of testing and the dangers we court when we ignore the principles and methods of educational measurement. Communication is a form of teaching and we should take the challenge of this kind of teaching more seriously than ever before. Perhaps we should consider using narratives as a framework for this effort. After all, using narratives (i.e., stories) to teach the elements of a subject has a long history and has been gaining popularity in fields as diverse as philosophy (Gaarder, 1995), process improvement (Goldratt & Cox, 1992), and test preparation (Marantz, 2003), among others. We have an obligation to be as creative in this effort as we pride ourselves on being in our technical research.

References

Blum, D. (2003). *Love at goon park: Harry Harlow and the science of affection*. Cambridge, MA: Perseus Publishing.

Braun, H. I. (2000). *A post-modern view of the problem of language assessment*. A. J. Kunnan (ed.) Cambridge, UK: Cambridge University Press.

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.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

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.

Gaarder, J. (1995). *Sophie's World: The Greek Philosophers*. Great Britain, UK: Phoenix House.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.

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

Goldratt, E. M., & Cox, J. (1992). *The goal: A process of ongoing improvement*. Great Barrington, MA: North River Press.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141-151.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29, 4-16.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, 32, 3-13.

Marantz, R. (2002). *The ring of McAllister*. New York: Simon & Schuster.

Mislevy, R.J. (in press). Substance and structure in assessment arguments. *Law, Probability, and Risk*.

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.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.

Sternberg, R.J. (1988). *The triarchic mind: A new theory of human intelligence*. New York, NY: Viking-Penguin.

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.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76, 105-110.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.

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.