

Heuristics and NCLB Standardized Tests: A Convenient Lie

Arnold Dodge*
Long Island University

Abstract

The No Child Left Behind Act of 2001 requires public schools in the United States to test students in grades 3-8. The author argues that this mandate has been supported by the public, in part, because of the “availability heuristic,” a phenomenon which occurs when people assess the probability of an event by the ease with which instances or occurrences can be brought to mind. These “mental short cuts,” which tend to oversimplify complex issues, are being employed by policy-makers in promoting standardized testing as the panacea for the problems of the public school system. The premises of this campaign include the “good intentions” to “leave no child behind,” the promise of improved accountability through high-stakes testing and the purported worthiness of test results. The author claims these premises are specious and examines their harmful potential for diverting resources, distracting educators and alarming children.

* Dr. Arnold Dodge is assistant professor and chair of the Department of Educational Leadership and Administration, Long Island University, C.W. Post Campus, Brookville, N.Y. He served in the public schools for 37 years as a teacher, principal, curriculum administrator and superintendent. He has chaired numerous high-stakes testing forums and has lectured in the public schools and at the university on the impact of high stakes testing on classroom and school environments. He is currently working on a project with the Deputy Director of the Health, Emotion and Behavior Laboratory at Yale University, researching sources of teacher stress and satisfaction.

Heuristics and NCLB Standardized Tests: A Convenient Lie

There is always an easy solution to every human problem – neat, plausible and wrong

~ H.L. Mencken

The film, “An Inconvenient Truth,” is so compelling because of the message it conveys. Global warming – arguably a “truth” – is hard to imagine. How can we accept a catastrophic scenario - despite the scientific evidence - as inevitable? We can’t fathom this apocalyptic vision - an inconvenient notion to say the least.

We have, in education circles today, a 180-degree turn of the global warming scenario - a convenient lie. We find it easy to believe that nationally mandated testing serves the public weal. How can we argue with the simple logic of testing students for accountability purposes? The approach appears to address our education woes. Appearances can be deceiving.

The No Child Left Behind Act (2001) which mandates that all public school students in grades 3-8 be tested in math and English (and most recently in science) produces a single score for each subject for each student in the country (Standards, assessment and accountability, 2008). Numbers on standardized tests seem to satisfy the public thirst for the simple and the chartable. No need to follow the messy and complicated developmental changes that children undergo nor, for that matter, attend to their creative, artistic and emotional growth, when there are standardized test scores which can be aggregated, disaggregated, archived and published on a graph in a newspaper. Many of us who have toiled in the public schools in the teaching and administrative ranks are nonplused at this turn of events. How could a literate and informed society become so smitten with such a limited measure of success for their schools and for their children? One answer may be found in the phenomenon known as heuristics.

Heuristics

Broadly, a heuristic can be defined as a mental “short cut.” Tversky and Kahneman (1974) may have been the first researchers to systematically examine this construct. They investigated how and why people rely on simplified operations to explain complex phenomena. While “heuristics” as an approach to explain things can be quite useful, it can also lead to “severe and systematic errors” (Tversky & Kahneman, 1974). The “availability heuristic” seems especially appropriate in its relationship to the public’s perceptions of standardized testing as a measure of school and student success.

The availability heuristic occurs when people assess the probability of an event by the ease with which instances or occurrences can be brought to mind (Tversky & Kahneman, 1974). The availability heuristic is “an oversimplified rule of thumb which occurs when people estimate the probability of an outcome based on how easy that outcome is to imagine. As such, vividly described, emotionally-charged possibilities will be perceived as being more likely than those that are harder to picture or are difficult to understand, resulting in a corresponding cognitive bias” (Economic Expert, n.d.). ChangingMinds.org, an Internet site devoted to

understanding “all aspects of how we change what others think, believe and feel,” offers this bit of advice on the utility value of the availability heuristic: “Make those things which you want the person to use for decision-making (perhaps at a later date) vivid and very easy to bring to mind, for example with repetition and visual language. Make those things that you do not want them to use, vague, abstract, complex or uncomfortable” (Changing Minds, n.d.).

The availability heuristic formula seems to be working on the public’s perception of our schools. In a paean to using the business model for schools, Hallinger and Snidvongs (2008) developed a laundry list of items that promote good customer relations in business, including relevance of products and services, pricing, customer loyalty, etc. They conclude that these concepts and practices are relevant to schools, especially in an “era of accountability” (Hallinger & Snidvongs, 2008). Rowan (1982) noted that the accountability of schools is fundamentally based upon the extent to which they satisfy the public’s perception of legitimacy. Here, then, is a prime example of how the availability heuristic shapes the logic of school improvement: If we can find criteria that the public perceive as legitimate, then we can use these criteria to measure the success of our schools. (Never mind that the criteria may not truly reflect improvement in learning. As long as the factors are perceived as legitimate, we have measures of accountability that will be accepted.)

Heuristics are woven into the fabric of the standardized testing milieu. The average citizen may be overwhelmed by the nuanced, organic, multi-faceted, and non-linear nature of a student’s educational development. To the rescue is a simpler and more convenient answer to fill the void. Politicians, the business community and the media encourage the trade off of complexity for simplicity so that school and student progress can be reduced to “understandable” numbers that appear “legitimate.” Those who advocate and support the one-size-fits-all testing mandated by the federal government call upon an array of strategies to support the simplified approach. Three premises which drive the public image of NCLB as a panacea for what ails the public schools are identified in this paper. Each relies and ultimately depends on the public’s needs for short cuts (i.e., heuristics) to understand school and student progress.

- NCLB is framed with the good intentions to “leave no child behind”
- Accountability is based on high-stakes testing
- Standardized tests yield results that matter

Each claim is specious when regarded in light of the deep, rich and supportive experiences children need for healthy development. What follows are examples of what happens when schools focus on standardized testing in an attempt to provide simple answers to complicated issues. In order to see through the haze of “heuristics and biases” (Tversky & Kahneman, 1974), I enlist the support of perspectives from the trenches and from those who have studied the developmental needs of youngsters.

Paving the Road with Good Intentions

There exist today a host of “good-intentioned” programs in the public schools that attempt to ready students for the rigors of testing. These initiatives are designed to *help students focus on academics*; what they appear to be doing is *getting students ready for tests*. It starts with kindergarten.

The original kindergarten (first established in 1837 in Germany) was created for children ages 3-7 years as a way to develop mentally, socially and emotionally through interactions with the outdoors and with opportunities for growth through movement, music and play. Friedrich Froebel, who coined the term and developed the first kindergarten class, based his program on the notion that “children need to have play time in order to learn. Kindergarten should be a place for children to grow and learn from their social interaction with other children” (Richie-Sharp, 2003, para. 2).

Kindergarten is no longer the “children’s garden” that was once envisioned. The focus recently has been on academics (Shepard, 1989), specifically reading readiness. Charts and graphs that detail letter and sound recognition growth, tests to determine spatial and temporal awareness, tests in math and reading (Gonen, 2008), and language experiences that deconstruct stories for literary elements are de rigueur for the kindergarten classroom. Kindergarten has become the academic farm team for the big leagues, aka first and second grade. In many schools recess has been reduced or eliminated for kindergartners (Nussbaum, 2006). This is a particularly ironic shift in that young children need “play” time to improve “think” time. Olga Jarrett, a professor at Georgia State University, has done extensive research on the importance of play and has found that on days that children had recess they were less fidgety and more on-task, with hyperactive children reaping the most benefit (Jarrett, 2002). To provide even more time for instruction, schools are lengthening the day for some kindergarten students. New York City Schools recently extended kindergarten hours for students who need extra help so that a typical school day can run over seven hours for these youngsters (Lucadamo, 2006).

Changes to the experience of the youngest denizens of the public schools are to ensure that no kindergartner is left behind. How can one argue with increased academic time in our schools? It seems so simple and well-intentioned. However, when five year olds are asked to put in overtime and when their play time is reduced, “good intentions” seem more like poor judgment. As Daniel Pink notes, we may be turning our young children into “automatots” (McCaw, 2007, p. 36).

As the curriculum gets more involved in the upper grades, the distortions continue. Reading education in some places takes a lethal dose of well-intentioned policy and practice. Teachers, pressured to increase reading scores to improve their schools’ NCLB profile, are spending inordinate amounts of time prepping for reading exams (Nichols & Berliner, 2008). At the same time, a study by the National Endowment for the Arts reported on a decline of daily pleasure reading among young people as they progress from elementary to high school. The decline appears to continue through college (Rich, 2007). The absurdity of conflating reading education with test prep is pointed out by a parent comment in a *New York Times* letter to the editor: “My son attends arguably the best public middle-school program in Baltimore, and the language arts teachers there have been told not to teach novels until the spring, after the state testing is over” (Myers, 2007). Another parent, on the same page, writes: “When classrooms are turned into test-preparation factories, reading scores may eventually rise, but those gains constitute a Pyrrhic victory because reading for pure enjoyment is destroyed” (Gardner, 2007).

In a stunning example of test prep undermining reading improvement, McNeil and Valenzuela report on the Texas version of NCLB accountability tests known as the Texas Assessment of Academic Skills (TAAS):

High school teachers report that although practice tests and classroom drills have raised the rate of passing for the reading section of the TAAS at their school, many of their students are unable to use those same skills for actual reading. These students are passing the TAAS reading section by being able to select among answers given. But they are not able to read assignments, to make meaning of literature, to complete reading assignments outside of class, or to connect reading assignments to other parts of the course such as discussion and writing. Middle school teachers report that the TAAS emphasis on reading short passages, then selecting answers to questions based on those short passages, has made it very difficult for students to handle a sustained reading assignment. After children spend several years in classes where “reading” assignments were increasingly TAAS practice materials, the middle school teachers in more than one district reported that (students) were unable to read a novel even two years below grade level. (as cited in Nichols & Berliner, 2007, p. 130)

A focus on reading and math test results – since this is where a district’s NCLB fortunes rise and fall – has wrought additional casualties in other disciplines. A Council for Basic Education study surveyed 956 elementary and secondary school principals in Illinois, Maryland, New York and New Mexico and found that there was a decreased emphasis on the arts and foreign language (Perkins-Gough, 2004). These subjects in many places seem to be regarded as vestigial, perhaps owing to the lack of value that NCLB assigns them. Hear the lament of a (former) elementary school teacher regarding mandated test prep and the disenfranchised subjects:

From my experience of being an elementary school teacher at a low-performing urban school in Los Angeles, I can say that the pressure became so intense that we had to show how every single lesson we taught connected to a standard that was going to be tested. This meant that art, music, and even science and social studies were not a priority and were hardly ever taught. We were forced to spend ninety percent of the instructional time on reading and math. This made teaching boring for me and was a huge part of why I decided to leave the profession. (as cited in Rothstein & Jacobsen, 2006)

Other promotions and initiatives to improve test scores are equally distressing – and sometimes expensive. In another putatively well-intentioned initiative under the NCLB banner, schools in which too many students fail math or reading exams must use federal funds to offer tutoring programs to low-income families. In the 2006-2007 school year, \$595 million went to the for-profit and non-profit tutoring industry. What are the results? Studies in Tennessee, Alabama, Georgia, Michigan and Kentucky showed that “supplemental educational services” did not improve test scores (Glod, 2008, para. 4). In a pay-for-performance plan, schools in New York City have adopted a plan to pay teachers and students who make improvements in test scores (Farley & Rosario, 2008). Preliminary results from the program are being reviewed by the State Education Department (Gonen & Soltis, 2008). The very notion of payment for improved test results – an idea that cynically offers the profit motive

as an available heuristic for public consumption – may be palatable to some business folks and politicians, but should be anathema to every committed educator and every parent who is concerned about instilling the love of learning in their children. And for those who need more evidence about the paucity of results from external rewards for learning, Nichols and Berliner (2008) opine:

A system of rewards, punishment, and pressures on self-esteem sounds like a logical way to motivate teachers and students, and some psychologists support this approach. But it doesn't work very well. Motivational researchers Richard Ryan and Kirk Brown present evidence strongly suggesting the opposite. They claim that it is the more autonomous motives, such as intrinsic motivation (e.g., I do it for *me*, not for *you*) or a well-internalized value system (e.g., I am guided by my own goals, not ones set by someone else), that result in higher quality of learning, persistence in the face of difficulty in learning, and greater enjoyment of the learning process. These are not the motivational systems elicited by high-stakes testing. (Nichols & Berliner, 2008, p. 149)

What has happened throughout the school systems of the United States, by and large, is that the voices of thoughtful educators who understand the richness of child development have been eclipsed by the hypnotizing drumbeat of those claiming to have a simpler answer: *if we can test each child, we can help each child*. This powerful short cut has hijacked the public's imagination. Those advocates of standardized tests hold up high-stakes accountability as the stick that is finally shaking up the educational establishment. The insistence that tests must be high stakes if they are to be worthwhile is another convenient lie that needs debunking.

High Stakes are for Gamblers

Competitive yoga. As foolish as the term sounds, it represents a movement to make yoga into a competitive sport. There is actually an organized group lobbying to make yoga into an Olympic event (ABC OF YOGA, 2006). The mentality that would drive a spiritual experience into a high-stakes competitive environment is the same mentality that thinks that a child's learning progress should be under the klieg lights while judges hold up signs with numbers. Therein resides another available short cut to fire the public imagination. Pressure to perform seems like an appropriate ethos within which to achieve optimal results from our students and teachers. After all, the conventional wisdom goes, when the going gets tough, the tough get going. And don't we all want to toughen up our schools to meet the demands of the 21st century?

The pressure to perform may suit those who voluntarily choose such venues, but to foist this arrangement onto a captive audience of youngsters is beyond the pale. High-stakes testing in the NCLB environment uses a threat of publicly announced failure to modify behavior. The former assistant secretary for elementary and secondary education during NCLB's inauguration weighs in on the "shame" factor: "The impetus for change built into NCLB was to effectively 'shame' schools into improvement. We now see that the shame game is flawed . . . The rhetoric of leaving no child behind has trumped reality" (Neuman, 2008, para. 5).

High-stakes testing - with its attendant threats and pressures - will not service us as an accurate accountability tool. Those who advocate for such an approach ignore the counterproductive effects of stress on performance.

Selye (1907-1982) pioneered research on the reaction of the body to stress. Selye's General Adaptation Syndrome outlines three stages of the body's adaptation to stress: ". . . an initial brief alarm reaction, followed by a prolonged period of resistance and a terminal stage of exhaustion and death" (Neylan, 1998, p.230). While citing these stages may seem overly dramatic in a discussion of reactions to school testing programs, there exists a similar trajectory that can be more easily applied to everyday stress and performance issues. The Yerkes-Dodson law (1908) provides a model to advance the conversation:

Arousal is a major aspect of many learning theories and is closely related to other concepts such as anxiety, attention, agitation, stress, and motivation. The arousal level can be thought of as how much capacity you have available to work with. One finding with respect to arousal is the Yerkes-Dodson law (1908) [which] predicts an inverted U-shaped function between arousal and performance. A certain amount of arousal can be a motivator toward change (with change in this discussion being learning). But too much or too little will certainly work against the learner. You want some mid-level of arousal to provide the motivation to change (learn). Too little arousal has an inert affect on the learner, while too much has a hyperactive affect. (Clark, 1999, para.1)

Goleman's description of the U-shaped curve is offered in a larger context of finding the "sweet spot" for optimal achievement:

An upside-down U graphs the relationship between levels of stress and mental performance such as learning or decision-making. Stress varies with challenge: at the low end, too little breeds disinterest and boredom, while as challenge increases it boosts interest, attention, and motivation – which at their optimal level produce maximum cognitive efficiency and achievement. As challenges continue to rise beyond our skill to handle them, stress intensifies; at its extreme, our performance and learning collapse. (Goleman, 2007, p. 271)

The climate surrounding the testing regime is highly charged and unforgiving – a breeding ground for intensifying stress. Students are primed for months before a test as if they were getting ready for battle. Reports of student anxiety are prevalent (Nichols & Berliner, 2008). But beyond the stress-laden climate, how does anxiety play into the significance of test results? A well-known critic of competition in schools, Kohn (2000), has studied the ill effects of pressures on children as they learn:

. . . test anxiety has grown into a subfield of educational psychology, and its prevalence means that the tests producing this reaction are not giving us a good picture of what many students really know and can do. The more a test is made to "count" – in terms of being the basis for promoting or retaining students, for funding or closing down schools – the more that anxiety is likely to rise and *the less valid the scores become*. (author's emphasis) (Kohn, 2000, p 5)

Stories about marching orders and pressures around test prep are legion. In many places the emphasis on test prep leaves a wake of missed opportunities. A teacher complains that the district's focus on reading, writing, and mathematics has precluded interesting experiences in hatching baby chicks or going on field trips or participating in community outreach (Rothstein and Jacobsen, 2006). A music director bemoans the fact that classroom teachers pressured to prep for tests no longer support the music program, some going to the extreme of sabotaging music lessons so that students do not leave the classroom. Teachers, this director says, will tell parents that music instruction "interferes with learning" (Seewald, 2007, p. 15). Some orders sound like triage protocols in an understaffed emergency room: A principal who told teachers ". . . to cross off the names of students who had virtually no chance of passing and those certain to pass. Those who remained, children on the cusp between success and failure, [should] receive 45 minutes of intensive test preparation four days a week, until further notice" (de Vise, 2007 para. 2).

It is perhaps easy to understand why the public is so enamored of high-stakes experiences. Sporting events, TV talent contests, food cooking competitions, etc. are the steady diet offered by the American popular culture. We want to be part of a winning team and we revere those who get the winning results. It is easy, then, to make the leap to want the same for our children. The simple proposition that high-stakes events lead to improved performance is another example of the availability heuristic at work. High-stakes tests and the results they yield are digestible information. What may turn the public's stomach, however, is an honest look at the tests themselves.

"The Mismeasure of Man"

In his seminal work, "The Mismeasure of Man," Stephen Jay Gould (1981) takes on the measurement community. In a wide-ranging assault on everything from craniometry to IQ tests, Gould lays out the argument that humans have a long and infamous history of mismeasuring one another. He stakes his claim on two fallacies: reification and rank. According to Gould, reification is

. . . our tendency to convert abstract concepts into entities (from the Latin *res*, or thing). We recognize the importance of mentality in our lives and wish to characterize it, in part so that we can make the divisions and distinctions among people that our cultural and political systems dictate. We therefore give the word 'intelligence' to this wondrously complex and multifaceted set of human capabilities. This shorthand symbol is then reified and intelligence achieves its dubious status as a unitary thing. (Gould, 1981, p. 24)

The second fallacy, ranking, Gould defines as

. . . our propensity for ordering complex variation as a gradual ascending scale . . . ranking requires a criterion for assigning all individuals to their proper status in the single series. And what better criterion than an objective number? Thus, the common style embodying both fallacies of thought has been quantification, or the measurement of intelligence as a single number for each person. (Gould, 1981, p. 24)

The designers of NCLB's high-stakes testing programs must have been channeling Gould when they thought up the idea of assigning numbers to student performance. Reification and ranking, as described by Gould, appear to fit conveniently under the availability heuristic umbrella. We can imagine a number and an order associated with our respective intelligence potential; anything more "wondrously" complex does not compute. How fitting then for policy makers to design a system that quantifies and ranks such ineffable and mysterious human skills as literacy and numeracy.

In a modern day version of "fool's gold" we believe that standardized testing is a system that gives us a data rich collection of student performance that accurately reflects each child's potential. With this mother lode of comparative statistics, we can evaluate and rank our students. Who would dare to question a state-sponsored regime that includes official uniform booklets for all students, directions for administration, guidelines for scoring, and score reports that quantify and order student performance? The program seems (a) efficient, (b) egalitarian, and most of all, (c) useful. In this author's opinion it is (d) none of the above.

On the issue of efficiency, there are examples of mismanaged administration and scoring throughout the country. Education Sector, a Washington-based think tank, surveyed 23 states in 2006 and found that 35% of testing offices had experienced "significant" errors in scoring and 20% didn't get results back "in a timely fashion" ("Testing Companies Struggle," 2007, para.6). The latter problem - not returning scores in a reasonable timeframe - is an egregious error in the effective use of test results. In New York, where grade 3-8 tests are administered to approximately 300,000 students a year, the English Language Arts test is given in January and the math test is given in March. Results are not scheduled for release until the end of the school year. As one Regent put it: "Is this information really valid for instructional planning when you take a test in January and get results six months later?" (Saunders, 2008, p. 4). Given the volatile nature of cognitive development in children through their early teen years (Elkind, 1981), test scores that are not returned for months are not only meaningless, but can be counterproductive. Scores on any exam have a shelf life; once expired, results that are used for diagnostic purposes can lead to poor instructional choices. Imagine receiving the results from a test for a medical condition months after the onset of the problem. By then the patient would have died, the condition worsened, or perhaps the more fortunate would have spontaneously recovered. Certainly, the nostrums that might have worked based on a timely diagnosis would be useless after the problem had run its course.

In his assessment of No Child Left Behind, Hursh (2008) uses New York's testing program as an example of NCLB-mandated test deficiencies:

. . . almost every recent standardized exam in New York has been criticized for having poorly constructed, misleading or erroneous questions or for using a grading scale that either over- or understates students' learning. Critics argue that an exam's degree of difficulty has varied depending on whether the State Education Department (SED) wants to increase the graduation rate (and therefore make the exams easier) or wants to appear rigorous and tough (and therefore makes the exam more difficult). (Hursh, 2008, p. 504)

Arenson suggests that incompetence may be a factor in high failure rates in some New York high school exam results:

Furthermore, sometimes an unusually low or high failure rate may not be intentional but the result of incompetence. The June 2003 Regents 'Math A' exam... was so poorly constructed that the test scores had to be discarded. Only 37% of the students passed statewide. (as cited in Hursh, 2008, p.505)

In an event sponsored by the National Academy of Sciences Committee on Incentives and Test-Based Accountability, a representative from the Educational Testing Service made these sweeping charges against NCLB tests:

. . . federally mandated education accountability systems [are] psychometrically weak, and predicated on mistrust between the actors and the system. We spend too much time . . . on outcomes, and not enough time on process, or collective human judgment. . . . he acknowledged that we had no idea what it meant, really, to be "proficient." In the absence of wisdom, we rely on single-number or composite-number metrics. (Flanagan, 2008, para. 7)

But what about the equality issue? Surely one can't quibble with a design that requires that the same measuring device be used for all children. In an odd version of noblesse oblige, those fortunate enough to have college educations, i.e., the policy makers, have designed a system, they believe, that will raise up those who have been educationally neglected by using the *same standard of measurement for all*. This notion seems eminently fair. Here is an argument that may be the mother of all oversimplifications.

The extraordinary differences in background, resources, and home environments that our students present to us each day across the country affect the way they perform in our schools. The skeptics can go to any school district's socio-economic status (SES) indicators and make a prediction regarding test score results. What they will find is that the correlation between SES and academic achievement is astonishingly strong (Hayes, 2004). The policy makers insist that we will leave no child behind if we test all children with the *same* instrument - a solution that fits the definition of an available heuristic quite neatly. What they don't focus on are the glaring inequities in home life that children bring to the schoolhouse and which ultimately affect their academic standing. All the high-stakes testing plans that can be mustered in state education departments and testing corporation headquarters, will not overcome the crushing effects of poverty in neighborhoods that are not equipped to support young, developing minds. (We should be cautious of temporary gains that are sometimes posted in inner-city elementary schools and hailed by NCLB advocates as signs that the testing juggernaut is working. These gains are often the result of "test-preparation regimens" and have little impact on secondary school performance) (Sanacore, 2007, p. 35). Neuman (2008), the erstwhile assistant secretary of education, is eloquent on the subject of poverty and schooling:

A child born poor will likely stay poor, likely live in an unsafe neighborhood, landscaped with little hope, with more security bars than quality day care or after school programs. This highly vulnerable community will have higher proportions of very young children, higher rates of single parenting, and fewer

educated adults. The child will likely find dilapidated schools, abandoned playgrounds, and teachers, though earnest, ready to throw in the towel. The child will drop further behind, with increasingly narrow options. (Neuman, 2008, para. 7)

As an ironic aside, many high-performing districts may be unfairly reaping the rewards of high-stakes testing results in the real estate sweepstakes. In a piece on the relationship between home values and test scores, the following is noted:

. . . overall test scores may reflect more on parental advantage than school quality. A student from a privileged background, in a high-income school district, may arrive at school well-prepared and start out scoring well on standardized tests. Years of schooling may not improve that student's scores. . . . On the other hand, a disadvantaged student in a different school district could end up improving his test scores more than the privileged student, all because he went to a high-quality school. But in the end, if his test scores are not as high as that of the privileged student, the school will not get as much credit, at least in terms of house prices. (Ascribe newswire, 2006, paras. 13,14)

Finally, the issue of the usefulness of the tests – i.e., are the test results giving us information that will help us to predict future success - is taken as an article of faith by an unwitting public. If it's a reading test, it must be useful in indicating how skillful students are in reading, and how they will perform in real-life situations when asked to read. Surely the tests must be valid instruments to guide us in our plans for the next generation. Read on.

Hear what Berliner and Nichols (2007) have to say about construct validity, the validity that tells us whether a test measures the abstract attribute or characteristic it claims to measure:

We found numerous examples from schools across the country that had dedicated hours upon hours preparing students for the test – drilling, emphasizing rote memorization, teaching students how to take tests, reviewing over and over again the concepts that will be represented on the test, and giving multiple practice tests, all at the expense of other content, ideas, and curricula that may not be represented on the test. At some point a line is crossed, and it messes up the interpretation of what a test score means. Construct validity is compromised when that line is crossed. No longer are we measuring real-world math or reading skills. Instead, it becomes a test of how well students memorized math content or how adept students are at filling in test-booklet bubbles. In these instances, it isn't content mastery that matters but how well (or efficiently) students can memorize information that is rewarded. (Nichols & Berliner, 2007, p. 122.)

Buttressing the test-prep/validity problem is the huge disparity that is being discovered between nationally administered NAEP exams and state administered NCLB tests. Michael Petrilli, a researcher at the Thomas B.

Fordham Institute in Washington poses this question in a report analyzing the differences between state and national tests: “The question is, why are the students making so much more progress on the state tests? What is likely to be happening is that schools are teaching students to that particular test” (Medina, 2007, p.B5).

Generally, the question of whether standardized tests measure what matters is troublesome. The real world calls for using knowledge in context, for the most part. Results from a measurement derived from an artificial testing environment will only tell us about how the test taker will do in an artificial testing environment, not how he or she will fare in the world, presumably the criterion that really matters.

Vygotsky, who studied how children learn and grow in groundbreaking work done during the early part of the 20th century, argued

. . . against standard intelligence and achievement testing procedures and against the view of development and education that emerges from the use of such tests. . . . He regarded the traditional tests of intellectual functioning of his time . . . as extremely limited because they only assessed “static” or “fossilized” abilities, leaving out the dynamic and ever-changing quality of human cognition (Berk & Winsler, 1995, p. 26)

Wineburg (1997) refers to the difference between Vygotsky’s approach and the more traditional view:

In contrast to traditional psychometric approaches, which seek to minimize variations in context to create uniform testing conditions, Vygotsky argues that human beings draw heavily on the specific features of their environment to structure and support mental activity. In other words, understanding how people think requires serious attention to the context in which their thought occurs. (Wineburg, 1997, p.4)

Perhaps most fundamental of all with regard to testing’s usefulness is whether what we are teaching is worth testing. With the emphasis on the tests themselves, there has been little time left to examine the curriculum. If we can believe Daniel Pink (2005), we are teaching ‘left brain’ skills to our children who are entering a ‘right brain’ world. Issues involving creativity, imagination, empathy, etc. are largely being ignored in the curriculum.

Routines and right answers are commodities. They are essentially free, anybody can do them, therefore they have zero or almost zero economic value. Whereas the ability to think, being able to be creative, to empathize with others, to tell a story, to listen to other people’s story; being adept at design, at connecting the dots, at recognizing patterns, at pursuing a life of purpose – those are not just the things that are going to enrich young people as human beings, but those are the types of things that our children are going to be doing for

a living. So there is a sort of a double whammy flaw in this routines and right answers obsession being used right now by many public school regimes. (Pink as cited in McCaw, 2007 pp. 35-36)

So while we trot out the ubiquitous comparative statistic tables that number and rank our children and our schools we become sanguine in our belief that the job is getting done. With a number and a rank, we are 'locked and loaded' with accountability information. No need to complicate matters with stories of test abnormalities or children's differing readiness to take on school tasks or whether or not the tests measure anything useful for the long term. Sir Kenneth Robinson, an internationally recognized author and lecturer on the subject of creativity, has this to say about the current state of public education:

Our education system has mined our minds in the way that we strip mine the earth for a particular commodity – and for the future it won't service . . . Our only hope for the future is to adopt a new conception of human ecology, one in which we start to reconstitute our conception of the richness of human capacity. (Robinson, 2007)

Snowflakes or Widgets?

What we have here is a failure to communicate. Those who believe that children need space and time and freedom to make mistakes, to exercise their imaginations as well as their bodies, to grow in fits and starts and on their own timetables, and to be understood as the complex organisms that they are, seem to be at odds with those who believe in packaging, promoting, distributing, codifying and simplifying school assessments. In short, some seem to believe that children are like snowflakes, unique and delicate. Others seem to believe that children are like widgets, uniform and shatter proof. The factory model approach is protected by those who claim to be offended by the "soft bigotry of low expectations" (Terkel, 2007). Like junkyard dogs, these barking voices protect the myth that shallow and often misleading data gleaned from one-size-fits-all testing can improve America's schools. While the public may "buy" these simplifications because they are available and appear reasonable, we may all need to take a collective breath and ask ourselves whether we are "buying" a convenient lie.

References

- Ascribe newswire (2006, July 12). *Student proficiency test scores impact home values, study finds*. Retrieved July 21, 2008, from <http://www.ascribe.org>
- ABC of Yoga (2006). *Competitive yoga to become an olympic sport*. Retrieved August 17, 2008, from <http://www.abc-of-yoga.com/news/200612/competitive-yoga-to-become-an-olympic-sport.asp>
- Arenson, K. (2003, August 27). New York math exam trial showed most students failing. *New York Times*, p. C12
- Berk, L.E., & Winsler, A. (1995). *Scaffolding children's learning: Vygotsky and early childhood education*. Washington, DC: National Association for the Education of Young Children.
- Changing Minds (n.d.). *Availability heuristic*. Retrieved July 26, 2008, from http://www.changingminds.org/explanations/theories/availability_heuristic.htm
- deVise, D. (2007, March 4). A concentrated approach to exams: Rockville school's efforts to raise questions of test-prep ethics. *Washington Post*. Retrieved March 6, 2007, from <http://www.washingtonpost.com>
- Economic Expert (n.d.). *Availability heuristic*. Retrieved July 26, 2008, from <http://www.economicexpert.com/a/Availability;heuristic.htm>
- Elkind, D. (1981). *The hurried child*. Reading, MA: Addison Wesley.
- Farley, J.D., & Rosario, H. (2008, May 15). A critique of monetary educational incentives for elementary and middle school students in New York City public schools. *Teachers College Record*. Retrieved May 24, 2008, from <http://www.tcrecord.org>
- Flanagan, N. (2007). Sitting in Einstein's lap. Retrieved January 15, 2008, from <http://teacherleaders.org>
- Glod, M. (2008, June 13). Mandated tutoring not helping Maryland, Virginia scores: Several states find 'No Child' provision does little to improve test results. *Washington Post*. Retrieved June 17, 2008, from <http://www.washingtonpost.com>
- Goleman, D. (2007). *Social intelligence: The new science of human relationships*. New York: Bantam.
- Gonen, Y. (2008, August 28). Parents slam 'k' kid testing. *New York Post*, p.2.
- Gonen, Y., & Soltis, A. (2008, June 24). School kids hit high marks. *New York Post*, p.8.

- Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.
- Hallinger, P. & Snidvongs, K. (2008, January). Educating leaders: Is there anything to learn from business management? *Educational Management, Administration and Leadership*, 36, 9-31
- Hayes, Deirdre (2004). State policy and district responsiveness: the relationship between district responsiveness to New York State policy in the areas of standards and assessments and district socio-economic status and district performance on assessment tests. (Doctoral dissertation, St. Johns University, 2004). *Dissertation Abstracts International*, 65, 2037
- Hursh, D. (2007). Assessing No Child Left Behind and the rise of neoliberal education policies. *American Educational Research Journal* 44(3), 493-518.
- Gardner, W. (2007, November 21). Reading for test scores and for life [Letter to the editor]. *New York Times*. P. A35
- Jarret, O. (2002). Recess in elementary school: What does the research say? Retrieved August 15, 2008, from <http://ceep.crc.uiuc.edu>
- Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth: Boynton/Cook.
- Lucadamo, K. (2006, February, 2). Hey, this is kindergarten, not salt mines, says irate dad. *New York Daily News*, p. 8.
- McCaw, D. (2007, February). Dangerous intersection ahead. *The School Administrator* 64 (2), 32-39
- Medina, J. (2007, November 16). Little progress for city schools on national test. *New York Times*. p. B5.
- Meyers, C. (2007, November 21). Reading for test scores and for life. [Letter to the editor]. *New York Times*, p. A35
- Neuman, Susan (2008, July31). Education should lift all children. *Detroit Free Press*. Retrieved August 15, 2008, from <http://freep.com/apps/pbcs.dll/frontpage>
- Neylan, T.C., (Ed.). (1998, spring). Hans Selye and the field of stress research. *Neuropsychiatry Classics*, 10 (1), 230
- Nichols, S., & Berliner, D. (2007). *Collateral damage : How high-stakes testing corrupts America's schools*. New York: Harvard Education Press.
- No Child Left Behind Act, PL107-110 (2001) (enacted).
- Nussbaum, D. (2006, December 10). Before the children ask, 'What's recess?' *New York Times*, p. 3.

- Perkins-Gough, D. (2004, September). Special report: The eroding curriculum. *Educational Leadership*, 62(1), 84-85
- Pink, D. (2006). *A whole new mind: Why right-brainers will rule the future*. New York: Riverhead Trade.
- Rich, M. (2007, November 19). Study links drop in test scores to a decline in time spent reading. *New York Times*. Retrieved November 27, 2007, from <http://www.nytimes.com>
- Richie-Sharp, S (n.d.). The history of kindergarten. Retrieved August 15, 2008, from <http://froebelweb.tripod.com/web2003.html>
- Robinson, K. (2006) Do schools kill creativity? Talk presented at TED Convention. Retrieved August 12, 2008, from http://www.ted.com/index.php/talks/ken_robinson_says_schools_kill_creativity.html
- Rothstein, R., & Jacobsen, R. (2006, December). The goals of education. *Phi Delta Kappan* 88(4), 264-272
- Rowan, B. (1982). Organizational structure and the institutional environment: The case of public schools. *Administrative Science Quarterly*, 27, 259-79.
- Sanacore, J. (2007). Needed: Critics of literacy education with a more inclusive perspective. *International Journal of Progressive Education*, 3(1), 29-43
- Saunders, Sylvia (2008, June 5) Regents seeking faster turnaround for 3-8 tests: Should SED revise testing calendar and change the scoring process? *New York Teacher*, p. 4
- Seewald, C. (2007). *Interviewing a school administrator*. Unpublished manuscript, Long Island University/C.W. Post Campus.
- Shepard, L., & Smith, M. (1989). Escalating kindergarten curriculum. Retrieved August 22, 2008, from <http://ericecece.org>
- Standards, Assessment and Accountability (2008). Retrieved August 22, 2008, from <http://www.ed.gov/admins/lead/account/saa.nt>
- Terkel, A. (2007). Bush's 'soft bigotry of low expectations' on Katrina reconstruction. Retrieved August 31, 2008, from <http://thinkprogress.org/2007/08/29/bush-education-Katrina/>
- Testing Companies Struggle to Meet NCLB Demands (2007, March 25). Retrieved March 29, 2007, from: <http://cnn.com/2007/EDUCATION/03/25.nclb.standardized.test.ap/index.html>
- Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131

Weingarten, R. (2007, October 18). Turning negative to positive. *New York Teacher/City Edition*. p.9

Wineburg, Sam (1997, September). T.S. Eliot, collaboration and the quandaries of assessment in a rapidly changing world. Retrieved January 21, 2008, from <http://www.pdkintl.org>

Yerkes-Dodson law: Arousal and learning (n.d.). Retrieved August 17, 2008, from <http://www.nwlink.com>