The Misfit Between Testing and Accountability

John Tanner, Test Sense

November 2016

In fall 2017, Texas will join 16 other states in implementing a public school rating system that assigns letter grades to schools and districts. By December 1, 2016, the Texas Education Agency (TEA) must adopt indicators showing how the A-F ratings will be determined, and by January 1, 2017, TEA must submit a report to the Texas House and Senate Education Committees showing the ratings that schools and districts would have been given if the system had been in place for the 2015–16 school year.

As we begin this important rule-making period, and as another Texas Legislature with authority to change the law that established Texas' A-F system prepares to meet, it is imperative that stakeholders know that the research is clear: A-F school rating systems fail as an indicator of school quality, but there is evidence that supports more meaningful kinds of accountability systems.

This essay is the third in the Texas Accountability Series, a series of essays that: provides an overview of A-F systems and their failures; explains why, to be meaningful, school accountability must be community-based and not solely focused on compliance with state testing mandates; and addresses the misfit of state testing programs with school accountability. (See also "The A-F Accountability Mistake" and "Creating a Meaningful Community-Based Accountability System.") Each of these essays was written by John Tanner, executive director of Test Sense and author of The Pitfalls of Reform.

As additional issues related to school accountability arise, the Texas Accountability Series will be continued to ensure that Texas educators have the information they need to work with policymakers and the public in a meaningful way.

Executive Summary

The argument: Rank order, standardized testing was invented to analyze human traits that could not be readily observed and for which no measuring stick existed. Their invention enables the rank ordering of a population on relative differences, and in turn allows an analysis to proceed in the absence of the measuring stick. The methodology never measured for the amount of anything.

Such tests work by finding a statistical average and then measuring out to the students furthest above and below average to create a ranking. The relative differences between students can then be observed and analyzed, even though a ranking can say nothing of what caused it to come to be.

Because such rankings are based on the aggregate of a student's experiences with the domain (e.g., numeracy or literacy), the patterns in the rankings will correlate with those experiences. If those experiences have patterns in society, then those patterns will be expressed in the rankings. Given that experiences with numeracy and literacy in the U.S. correlate highly with socioeconomic status, it is not surprising that the rankings do as well.

Ranking is one means by which the patterns in education can be viewed and disrupted. However, rankings are all too often assigned value judgments prior to knowing the reasons why a ranking is as it is. This is always a mistake: The causes for a ranking need to be determined first. Only then will it become clear if a judgment is warranted, or what types of changes should be supported.

Policymakers noticed that schools they perceived as good had high standardized test scores and declared that all schools should have high standardized test scores. The

Preferred Citation: Tanner, J., (2016). The Misfit Between Testing and Accountability. The Texas Accountability Series. Austin, TX: The Texas Association of School Administrators.

impossibility of this notwithstanding (e.g., it is impossible for everyone to be above average), policymakers have shown little interest in understanding the realities behind their primary educational accountability instrument. The State of Texas Assessments of Academic Readiness program (STAAR) in Texas is based upon rank order, standardized test methodologies.

Conclusion: A methodology designed to show the rank ordering of a population automatically sacrifices any capacity to comment on what caused that ranking. It can serve only as a signal for researchers to begin their search. The quality determination of schools and the passing score for students has, for the duration of education reform, been made via an instrument stripped of any ability to judge quality. This represents a grave concern, as the consequences are extensive.

Standardized Testing[1] as a High-Stakes Accountability Tool

At the advent of the school accountability era the notion that schools should do better was entirely reasonable. Schools struggled to serve all students well, with educational outcomes mirroring broader socioeconomic conditions that schooling could help address. Overcoming these challenges was paramount: Students in a democracy have the right to a high-quality education, and the preservation and growth of the U.S. economy depended (and continues to depend) on a great many more of our students achieving excellence than the number of them who had done so historically.

High or rising standardized test scores as a quality measure entered education in an extremely simplistic manner: Students and schools perceived as being good scored high on such tests, and schools perceived as being bad scored low. Therefore, all schools should score high, or absent that, demonstrate that their scores are improving. In recent years, a variety of federal requirements and sanctions were imposed should the perceived benefit of high or rising test scores not occur. While recent changes in federal policy eased the requirement for sanctions, the testing requirement remains, effectively shifting the decision regarding rewards and sanctions to the states. Each state must now decide how best to handle things going forward.[2]

"It is a surprise to many when they learn that the primary function for the tests used by states for accountability is to rank order students, not assign judgments of quality. In fact, the manner in which such tests perform their rank ordering precludes the resulting ranking from ever signaling quality"

Since the beginning of the accountability era the capacity of high or rising test scores to measure quality has gone largely unquestioned from both a federal and a state policy perspective. It can reasonably be said that the policy community and the public at large have long believed and continue to believe that standardized test scores are a reasonable tool for judging the quality of schools and dictating the application of consequences.

It is a surprise to many when they learn that the primary function for the tests used by states for accountability is to rank order students, not assign judgments of quality. In fact, the manner in which such tests perform their rank ordering precludes the resulting ranking from ever signaling quality. Nevertheless, policymakers and the public at large bought into the fallacy that a high ranking signals goodness while a low ranking signals just the opposite.[3] That belief creates a picture of school quality that is skewed at best, and more often than not, dead wrong.

Why Rank?

Ranking is just one tool among many used to study the characteristics within a population. When human beings can be ranked in terms of characteristics or behaviors—birth weight, baseball players' batting averages, or salaries, for example—it becomes possible to detect patterns in the ranking. Such patterns can serve as a signal to act. When those patterns reflect unfairness, an inequity, or something

deemed worthy of a change, they can provide an incentive to attempt a remedy. Additional rankings performed at a later date can help determine if the actions had the desired effect.

"The greatest challenge with a ranking is that it is so easily over-interpreted."

The greatest challenge with a ranking is that it is so easily over-interpreted. Rankings begin with the person with the least of the thing being analyzed and then proceed to the person with the most of the thing. The higher one is in such a ranking, the greater the practical advantage one also likely has, but it is a mistake to equate advantage with worth: When a study shows that men earn a higher salary for the same work as women, for example, it is not because a man has greater worth than a woman, but because something in society has created that advantage. The fact that being a man comes with certain advantages in this regard, just as being a woman comes with certain disadvantages, signals nothing about the worth of either. Advantage is a societal effect that a rank ordering reveals.

"…it is a mistake to equate advantage with worth: when a study shows that men earn a higher salary for the same work as women, for example, it is not because a man has greater worth than a woman, but because something in society has created that advantage…Advantage is a societal effect that a rank ordering reveals."

Ranking students in terms of educational attainment should, theoretically at least, work in a similar fashion. Such a ranking would show patterns, such as the impact of poverty, or gender inequality, which would be worthy targets for actions that attempt to disrupt them. Ranking would certainly not be the only way one should attempt to understand and analyze education, as it would provide a partial view at best, but it would nevertheless be useful.

Ranking Things Without a Measuring Tape

Ranking educational attainment offers a unique set of challenges. The first is that educators lack a measuring tape for things such as numeracy and literacy. Ranking babies on birth weight requires a scale, and ranking people on their salaries requires having their actual salaries. But what if neither was available. Could you still rank them? You could if you could observe the relative differences between them.

Height offers a clear example of how this could occur. Imagine how simple it would be to ask a room full of people to line up from the shortest to the tallest; the process takes minutes at most. Note as well that they can perform this ranking without a measuring tape, the thing they would need had the question been about how tall each person actually is.

Once that ranking is in place, no other information is needed in order to perform a range of statistical calculations that enable all sorts of analysis. In the middle of the ranking is the median position, which is one way of understanding average,[4] and from that position of average it is possible to further parse the population to help create nuances that can aid in interpretation.

It then becomes possible not just to identify patterns (e.g., men will be taller than women), but to compare populations. If people in the next room also perform a rank order on height you could march them in to see how they compare. The people from the second room would find the person in the original room that corresponded to their height and stand across from him or her. If you put a hat on the person at each average position, you could easily observe any differences in averages. If you parsed each ranking into chunks (e.g., how the top 10% in each compared), the observations become even more nuanced.

If you have the chance to perform that original ranking on a nationally representative group, the comparisons are even more useful, since now you are not just comparing two rooms of people, but rather the comparison is to a group that can stand in for the entire population. Since the average

position across the entire population will tend to change only slightly across time, researchers can create a stable tool for analysis both as of the moment of the original measure, and across time. All they have to do in order to create such a tool is observe the relative differences across a population. Having a measuring tape would be nice, but it is not necessary.

What if There Are No Relative Differences to Observe?

Unlike height, observing the relative differences between students in numeracy and literacy offers a bit of a challenge. A room full of random students asked to rank order themselves accordingly would be unable to do so.

To rank students, what is needed is a way to create observations of the relative differences in numeracy and literacy. Doing so is actually fairly simple, but it requires a different starting point than in the height example: Rather than start with the observations and then later find average, it is actually possible to find average and then work out from that point.

That process begins with a test question that half of the population will answer correctly and half will answer incorrectly. Once the population is divided according to their responses, the line between them can be said to represent a sort of average, representing the midpoint in a two-step ranking. But ranking into just an above-average and a below-average group isn't particularly useful for the purpose of analysis, so a second question is asked that like the first one also is known to divide the population in two. The second question creates three ranked steps: those who answered both of them wrong, those who answered one right and one wrong, and those who answered both of them right.[5]

The third question will also need to divide the population in two, but now an additional requirement for the questions is needed: Those at the top of the rankings will, for the most part, need to answer each question correctly, while those at the bottom will, for the most part, need to answer each question incorrectly. Because the point is to create observable differences splitting the population cannot just be a version of a coin toss. If it were, then fifty questions would result in lots of students with twenty-five correct and twenty-five incorrect responses, which create observations that are too similar to be useful, since they fail to show differences. Each question needs to divide the population similarly so that the test takers can be ranked from the student furthest below average to the student furthest above average.

For the same reason, no questions would be included that all students will answer correctly or incorrectly. Such questions may be important from an instructional or other perspective, but as they would show all students as similar, they are useless in creating observable differences. If they were included they would represent a waste of time and resources.

In the end, what can be produced from the test that emerges is a rank ordering of students based on relative differences. That ranking was produced without a measuring tape, so no amount of numeracy or literacy can be known at any particular point. The result is simply an estimate of a student's current ability compared to both the average and others. Rather than suggest that an estimate is "above average" or "below average" or somewhere in between, it is more convenient to assign numbers to each point in the range. The numbers don't much matter: you just need to pick a starting point, assign it average and work out from that point. That is why so many tests of this variety start with a number in the hundreds: it gives them plenty of room in both directions.

Once You Have Your Test…

The first large-scale educational tests to follow the rank order formula were norm-referenced tests offered by commercial vendors. The Stanford Achievement Test, The Iowa Test of Basic Skills, and the California Achievement Test are/were all examples of the methodology.[6]

"Interpretation of any rankings…requires extreme caution, since the interpretations available within a ranking are limited to analysis, and should never veer into judgments absent other information."

The tests rank ordered test takers according to the acquisition of numeracy or literacy lifetime to date. The publishers also *normed* the tests after they were created in an effort to make the comparisons and the patterns in the rankings more useful. Norming studies enable comparisons to be made from one grade to the next, as well as to compare the relative differences between various domains being tested. In that way researchers could identify and follow patterns as of a particular moment and across time in a very nuanced manner.

As but one example, a norming study will allow for adjustments to be made in the numbering system such that ten points of difference in one part of the scale is equivalent to ten points of difference in another. Absent the norming study any claim that a ten-point movement in one part of the scale represents a similar effort to ten point gains in other parts of the scale cannot be made. Or it would allow for educators to compare performance across domains, which cannot be done absent such a study.

Interpretation of any rankings—normed or otherwise—requires extreme caution, since the interpretations available within a ranking are limited to analysis and should never veer into judgments absent other information. In that regard it falls into the exact same category as a rank ordering of differences in height. It would be absurd to presume that a position in the ranking or a particular pattern regarding height comes with a known cause. People rank where they do in terms of height for any number of causes, including race, genetics, nutrition, age, and gender. That other information is what requires interpretations that lead to judgments. Changes in the ranking or the patterns in a ranking merely serve as a signal for where to look.

Norm-referenced test scores in education have never enjoyed that sort of objectivity even though the design requires it for proper interpretation of the results. Rankings of students have never escaped the bias that the ranking signals something about the student's nature as opposed to the aggregate of conditions and experiences that have made up the student's life to that point. It is those conditions and experiences that require interpretation and, where necessary, judgment and actions. The ranking is designed to explore those patterns and their causes, but certainly not judge them.

It is equally critical to remember that no rank order test, normed or otherwise, was designed to measure the amount of the thing being analyzed. Rather, they were designed to rank order students based on observable relative differences so that the patterns revealed in the patterns in those rankings could be analyzed and, where appropriate, actions and judgments could be taken.

"It is…critical to remember that no rank order test…was designed to measure the amount of the thing being analyzed."

All State Tests Follow the Rank Order Formula

Some classroom educators dislike norm-referenced testing for two reasons: They object to a test based on differences between students, and they are instructionally useless.[7] When the notion of educational reform first surfaced thirty years ago, so too did the notion that a new type of testing would be needed. The idea was that a system based on excellence for all would need to be measured by something other than a test designed to sort and rank students.

However, what those responsible for reform quite literally did was to adopt the rank order design, absent the norming studies. Every single component required to produce a rank ordering was preserved, but the resulting instruments were renamed things such as "criterion-referenced" or "standards-based" tests, as if a new name on an old methodology could magically transform it into something new.

Somewhere along the line a mistake was made in thinking that the norms created by the publishers created the comparisons of differences between students, and that absent those norms the comparisons would be to something else. Whoever made that mistake should have known norms are something applied to a rank ordering to aid in interpretation, and that eliminating them simply made the rank-ordered estimates of student achievement trickier to compare and interpret. Policymakers and reformers wanted to stop comparing students to students and instead start comparing them to an expectation, but they did not wish to give up the stability in the old scores that enabled comparisons over time.

"…what those responsible for reform quite literally did was to adopt the rank order design, absent the norming studies."

In other words, they wanted some of what the old norm-referenced systems possessed and figured that simply by jettisoning the norms what would be left were the parts they considered useful. The problem was that the parts they wanted to keep were all a part of the initial rank order design rather than the norms that enhanced the ability to perform analyses. Removing the norms merely removed that enhanced ability but left the rank ordering components intact.

Rankings Require Other Information

A rank order based on relative differences is very poorly suited for doing much more than being a rank order, but as a rank order it is useful for the comparisons it enables. One could, for example, compare the rankings of two students with similar experiences to determine if a difference exists. If it does, if one student ranks higher than the other, that can serve as a signal to begin the search for causes. The difference may exist as a result of decisions made by a school or a parent, maturity, a traumatic event in one child's life, cheating, or an excellent teacher, to name but a few.

A misinterpretation occurs should that difference trigger any judgment prior to determining why the difference exists. As with any premature judgment, it would be both inappropriate and highly likely to be wrong. Given the number of potential causes behind such a difference it would be common sense to realize that *maybe* something good or bad happened to create that difference, but *maybe* it did not.

"It makes no sense under any conditions to offer a reward or inflict a punishment without first knowing if either was warranted, and yet that is exactly what happens if a difference in rankings is assigned a judgment absent other information."

It makes no sense under any conditions to offer a reward or inflict a punishment without first knowing if either was warranted, and yet that is exactly what happens if a difference in rankings is assigned a judgment absent other information. Good decisions risk being punished, bad ones risk being rewarded, and things that no one can or should take credit for would risk being the basis for an award or sanction. In a system where no one bothered to look a correct judgment could only be attributed to luck, making it difficult to repeat positive practices or jettison those that aren't working.

The establishment of a cut score (or passing score) on such tests represents a similar misunderstanding. Drawing such a line is supposed to signal quality if the student is above it, and a lack of quality if the student is below it, turning the instrument into a machine that makes lots of judgments absent any evidence for doing so. A ranked position can never automatically signal what causes students to rank at that point absent a search for answers. The reasons will always vary and to presume anything about the answers prior to a search should be seen as unreasonable.

Declaring a judgment prior to having the evidence is dangerous in that it risks adding a bias during any subsequent searches for answers: A good judgment will start a search for validation points, while a bad judgment will start a search for points of failure. This bias is wrong in that it can be deeply

debilitating to schools and the students they serve. A high scoring school will focus on what should be repeated to preserve those high scores, which takes the focus away from patterns that could be disrupted in the name of improvement. When a low scoring school is perceived as failing, *any* change is perceived as acceptable, and often equally valid with any other changes, absent any evidence that the change is appropriate, since at the very least it will be *different*.

All schools need to constantly change and adapt, and no school should make wholesale change absent the right information to guide the change. A failure to change, or a failure to change the right things, is a recipe for stagnation and inefficiency.

Rankings Mean Someone Always Has to be at the Bottom

Assigning a point along a rank ordering the meaning of pass/no pass runs counter to the idea that the educational system should be about promoting excellence for all. Any passing score placed anywhere on a ranking nullifies that idea, since not all students will ever move past any point on a ranking—it would then no longer be a ranking.

Since the majority of passing scores are drawn around the average score, in order for all students to succeed schools will need to figure out how to help all students rank above average, which is utterly illogical.[8] The thing all students are supposed to strive for and achieve is in fact a point in a ranking that ensures only some of them will get there.

"The thing all students are supposed to strive for and achieve is in fact a point in a ranking that ensures only some of them will get there."

The Contributions of Schooling to Rankings

The assignment of an accountability label to a school based on where students rank on a test presumes that the school caused those rankings to come to be, and yet where a student ranks as of the testing date can accurately be expressed as an amalgam of a student's numeracy and literacy experiences to date.[9] That includes the school year in which testing occurred and all the schooling years leading up to that point, as well as all a host of non-schooling factors.

"…if the goal of accountability is to determine that teachers taught effectively and that students learned what was taught, or to understand the quality of a school, a ranking is the wrong tool with which to make that determination."

Numeracy and literacy are such that even when the content for a test is derived from the assigned grade, that content represents an additional layer upon principles and ideas that were first presented in prior years. A single year of learning is anything but a vacuum, but instead one more layer of a highly complex, iterative process.

Researchers place a surprisingly (to most) small percentage of what influences a student's ranking via a test score on schooling,[10] suggesting that about two-thirds of the effect represented in a ranking should be attributed to non-schooling factors. Researchers report a wide variance in terms of how much of the change over time regarding where a student ranks (i.e., "gains") can be attributed to a school or classroom (forty to seventy percent, depending upon the analysis being performed), determined by a host of factors, including the random assignment of students to teachers, the inability to actually attribute learning to a particular cause,[11] and the wide variety of other factors that frequently interact in non-predictable ways to influence achievement.

In order for a ranking based on the past to provide useful patterns that can influence what might be done in the future it needs to include the cumulative effect to date, regardless of where that effect comes from. It would be useless (and impossible) to limit the ranking or a change in the ranking to only those things schools control, since the patterns and comparisons of the past need to include all of

the literacy and numeracy ability acquired lifetime to date. Only then can a ranking be useful, since it is a school's job to disrupt those broader patterns going forward, regardless of their cause.

As a result, if the goal of accountability is to determine that teachers taught effectively and that students learned what was taught, or to understand the quality of a school, a ranking is the wrong tool with which to make that determination. To ask a ranking of numeracy or literacy, acquired lifetime to date, to carry accountability, is to place a non-trivial part of accountability directly on things the school or the past year's teacher do not control.

"Unless schools are allowed to investigate causes for a ranking, they risk getting credit for things they did not do, being blamed for things they had nothing to do with, and missing out on the reward for having made good decisions in the best interests of the students they serve."

Schools and the students each serves deserve to understand the quality of the decisions being made, which a ranking on its own can never signal. The reason for this is simple: A ranking cannot on its own unravel the various effects that caused it to come to be.

The ranking, or a change in the ranking, can signal that something happened, but to act as if the cause can be known the instant the ranking is known is akin to performing magic, because at that moment any and all causes are guesses.

Unless schools are allowed to investigate causes for a ranking, they risk getting credit for things they did not do, being blamed for things they had nothing to do with, and missing out on the reward for having made good decisions in the best interests of the students they serve.

So What if It Doesn't Make Sense?

Rank ordering students in terms of relative differences in numeracy and literacy makes sense if the goal is to find patterns that need to be disrupted. But assigning value to schools and students based on where each ranks does not.

Where that leaves education in the era of test-based accountability is in a terrific bind: When accountability is placed in a test designed to rank students, the practical result will be to negatively judge schools that serve students who have had fewer opportunities than their more advantaged peers, while positively judging schools with the more advantaged students for the mere fact that they are more advantaged.[12] That means teachers serving disadvantaged students are likely to be designated as bad teachers, while teachers serving advantaged students are likely to be designated as good teachers, and no one will know the actual truth.

"That means teachers serving disadvantaged students are likely to be designated as bad teachers, while teachers serving advantaged students are likely to be designated as good teachers, and no one will know the actual truth."

Policymakers wanted the best of several worlds. They wanted:

1. the statistical elegance of standardized test scores that were consistent, and therefore believable, over time
2. a measure that signaled excellence, and that could be used to judge the quality of schools
3. a measure on which all students could be successful
4. a way to evaluate the quality of teaching
5. a measure that would signal real proficiency for students

At this point educational policy is one for five: The system is statistically elegant, but at the expense of everything else. They managed to turn a methodology capable of analyzing differences in

educational opportunity into a means for punishing schools in which the students have not yet had the opportunities of their more privileged peers. They selected a testing methodology that reflects a lifetime of practice, rather than one limited to the past school year.[13] They selected a methodology that was never designed to judge quality, nor signal proficiency, and ask it to do all these things and more.

Conclusion

Rank ordering a population based on test scores offers researchers a useful way to study the patterns in education. Because statistical averages are by definition fairly stable over time, changes in average, or in a ranked individual's relative position to average, offer a clear signal that something happened to cause that change and is then worthy of a closer look as to potential causes for the change. Patterns can be identified and tracked over time, and such tools, in concert with others, can help support a continuous improvement paradigm.

"At this point educational policy is one for five: The system is statistically elegant, but at the expense of everything else."

"The worst part about the improper use of standardized test scores in an accountability system is that it leaves policymakers and the public in the dark regarding the one thing they can and should care about more than all others: the actual quality of our schools and what can be done to improve."

Policymakers rightly believed that they and the public need information as to how schools are performing, school quality, and how to motivate bad schools to improve. They then selected a tool designed for another purpose entirely, one that can tell them where a school or student ranks, but cannot tell them why or what caused that ranking.

Absent all the critical pieces of information that should be required before judging a school or a student, they nevertheless presume to have the answer: High-scoring schools and students reflect quality, while low-scoring schools and students do not. Rankings are supposed to be objective so that the patterns in them can be explored and, when necessary, disrupted. Using rankings as a tool of judgment is a means to stagnation and inefficiency, effectively preserving the patterns the tool was supposed to help disrupt.

The worst part about the improper use of standardized test scores in an accountability system is that it leaves policymakers and the public in the dark regarding the one thing they can and should care about more than all others: the actual quality of our schools and what can be done to improve.

Notes

[1] The term "standardized" actually refers to the conditions under which a test can be administered. It means that all test takers are to approach the test under the same conditions. Standardization is particularly important when a test is designed to compare test takers and the differences between them. While any test, quiz or assessment can be standardized the phrase "standardized testing" is now ubiquitous for referencing the methodology that underlies all state testing programs, including STAAR here in Texas. It is in that vein that the term is invoked here.

[2] More information on the Every Student Succeeds Act can be found at http://www.aasa.org/AASAESSA.aspx

[3] U.S. Congress, Office of Technology Assessment, (1992). "Lessons From the Past: A History of Educational Testing in the United States, Testing in American Schools: Asking the Right Questions.

Washington, DC: U.S. Government Printing Office. Retrieved from http://govinfo.library.unt.edu/ota/Ota_1/DATA/1992/9236.PDF

[4] Statisticians use three different versions of average depending on which is appropriate for the data being analyzed: mean, in which all of the results are added and then divided by the number of measures; mode, which is the most common result; and median, which is the mid point in the available records. All three help to understand the central tendency in a data set. In the case of analysis where no measuring stick is available, the median and the mode are still frequently available, since they can be calculated independent of the data required for a mean.

[5] For the sake of brevity I have simplified my explanation regarding the manner in which such tests are built. I ask for the indulgence of measurement professionals in attempting to make what is a highly complex activity understandable.

[6] For those seeking additional information on the efforts of psychologists a century ago who were the first to use tests to rank order for intelligence tests, see Gould, Steven Jay, *The Mismeasure of Man* (New York: W. W. Norton and Company, 1996). While testing for intelligence is different than educational testing that also rank orders students, the underlying desire to rank is the same. Seeing the history of how and why that occurred is enlightening.

[7] Popham, W. J, "Why Standardized Tests Don't Measure Educational Quality," *Educational Leadership*, (6) 56, (1999), 8-15

[8] See, for example, http://tea.texas.gov/Student_Testing_and_Accountability/Testing/, which includes frequency distributions for each STAAR test given, as well as overall results. What quickly becomes obvious after reviewing a dozen or so tests is how many of them set the current cut score around the median mark in terms of students, and that average seems to be a frequent target. In the case of the future targets, however, many of the cut scores are significantly above current averages.

[9] Rowna, B., Correnti, R., and Miller, R., (2002). What Large-Scale, Survey Research Tells Us About Teacher Effects On Student Achievement: Insights from the Prospects Study of Elementary School. Philadelphia, PA, Consortium for Policy Research in Education, University of Pennsylvania.

[10] Rowna, B., et. al.

[11] Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R., & Shepard, L., (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. Washington, DC: The Economic Policy Institute.

[12] Adams, C. M., Forsyth, P. B., Ware, J. K., Mwavita, M., Barnes, L., & Khojasteh, J. An empirical test of Oklahoma's A-F grades. Education Policy Analysis Archives, 24(4). http://dx.doi.org/10.14507/epaa.v24.2127

[13] Here an explanation may help ward off argument that in the current system the test content is selected from that year's educational content, and therefore the test is only about the current year's material. Two things are worth mentioning in this regard. First, subjects like numeracy and literacy take a lifetime to acquire, with each step building upon the next. Learning the current content requires all of the experiences to date that have led to the capacity to learn it at that point. Second, standardized test items are selected for their ability to discriminate between above and below average. Every item on STAAR and every state test used for accountability follows this pattern. Since learning takes place over a lifetime, and since the discrimination of an item represents what has been learned over a lifetime, it cannot be said that selecting the content from this year's list makes the test only about what is on the test. In fact, nothing could be further from the truth.