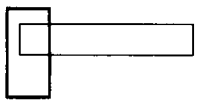

# The Problem With "Proficiency": Limitations of Statistics and Policy Under No Child Left Behind

Andrew Dean Ho

The Percentage of Proficient Students (PPS) has become a ubiquitous statistic under the No Child Left Behind Act. This focus on proficiency has statistical and substantive costs. The author demonstrates that the PPS metric offers only limited and unrepresentative depictions of large-scale test score trends, gaps, and gap trends. The limitations are unpredictable, dramatic, and difficult to correct in the absence of other data. Interpretation of these depictions generally leads to incorrect or incomplete inferences about distributional change. The author shows how the statistical shortcomings of these depictions extend to shortcomings of policy, from exclusively encouraging score gains near the proficiency cut score to shortsighted comparisons of state and national testing results. The author proposes alternatives for large-scale score reporting and argues that a distribution-wide perspective on results is required for any serious analysis of test score data, including "growth"-related results under the recent Growth Model Pilot Program.

**Keywords:** accountability; high-stakes testing; statistics

In John Godfrey Saxe's famous poem *The Blind Men and the Elephant,* six men, all of whom are blind, disagree about what they are observing. One man argues that he is touching a wall; another, a snake; and the others, a spear, a tree trunk, a fan, and a rope, respectively. The moral of the story is straightforward. All of the men are in a sense correct in that they are accurately describing portions of a larger whole: an elephant's side, trunk, tusk, and so on. They are also in a sense incorrect in that they are missing the so-called big picture. They do not recognize that they are all in fact observing a single larger entity.

The Percentage of Proficient Students (PPS) is a conceptually simple score-reporting metric that became widely used under the National Assessment of Educational Progress (NAEP) in the 1990s (Rothstein, Jacobsen, & Wilder, 2006). Since 2001, PPS has been the primary metric for school accountability decisions under the No Child Left Behind (NCLB) Act. In this article, through a hierarchical argument, I demonstrate that the idea of proficiency—although benign as it represents a goal—encourages higher order interpretations about the progress of students and schools that are limiting and often inaccurate. I show that overreliance on proficiency as a reporting metric leads to statistics and policy responses that are overly sensitive to students near the proficiency cut score, just as Saxe's blind men can describe only what is immediately in front of them.

My argument against overuse of the PPS statistic progresses through five levels (Table 1). Level 1 is a reminder that PPS statistics are dependent on a cut score judgmentally determined by a standard-setting process (Hambleton & Pitoniak, 2006). Reasonable disagreements may arise about where this cut score should be located, and there is no wholly objective method of determining a "true" cut score (Linn, 2003). Level 1 concerns about PPS statistics are becoming widespread (e.g., McCabe, 2006; Wallis & Steptoe, 2007), particularly where two different tests report vastly different PPS for the same state. Interesting analyses have been conducted by Braun and Qian (2007) and McLaughlin and Bandeira de Mello (2005), who use mapping techniques to quantify the differences between state and NAEP proficiency cut scores. At Level 1, the degree of proficiency is framed as a choice. At higher levels, this choice is seen to affect interpretations of progress, equity, and policy in dramatic ways.

At Level 2, I demonstrate that the magnitudes of PPS-based trends and gaps depend on the selection of a cut score. A good demonstration of Level 2 concerns is a simple figure (Figure 1) showing PPS trajectories for six hypothetical states in the NCLB era. At first glance, it is apparent that all six states are improving over time, and the states' rankings on their percentages of proficient students is their alphabetical order. Further inspection reveals that rates of improvement vary among states. States B and C, for example, make considerable gains early in the NCLB era, increasing around 10 percentage points between 2006 and 2007. Over time, however, the rates of improvement for States B and C decline. Meanwhile, State F, which started off with marginal gains, begins to make much greater gains in the middle years. These comparisons look substantively meaningful, and one might imagine that these observations would make headlines or inspire a research report.

The punch line is that these states have the same data. Their test score distributions are identical, and the curves drawn in Figure 1 are different only because of each state's choice of proficiency cut score. The range of their starting points is a Level 1 concern. The differences among their trajectories are a Level 2 concern. Contrasting trajectories across these states is akin to generalizing from unrepresentative fragments and disagreeing about what is in fact a singular whole. Figures or tables that track the

Educational Researcher, Vol. 37, No. 6, pp. 351–360
DOI: 10.3102/0013189X08323842

**Table 1**
***Hierarchical Properties of Proficiency-Based Statistics***

| | |
|---|---|
| Level 1 | Proficiency cut scores are judgmental. |
| Level 2 | Trend and gap magnitudes depend on proficiency cut scores. |
| Level 3 | Gap trend magnitudes depend dramatically on proficiency cut scores. |
| Level 4 | These dependencies are not straightforward. |
| Level 5 | These dependencies may be used, cynically or nobly, as policy tools. |

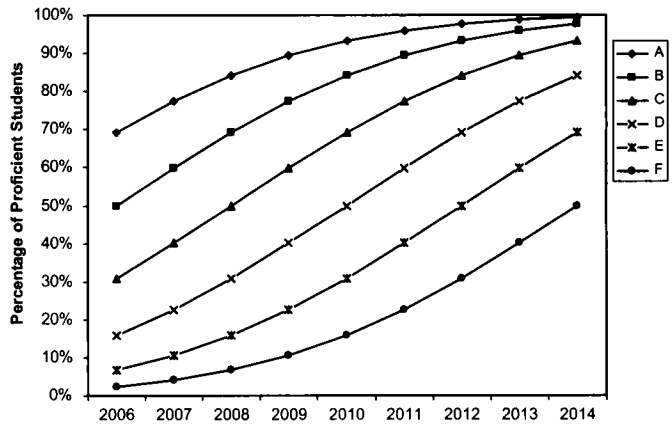


FIGURE 1. *Percentage of Proficient Students trajectories for six "blind" states.*

percentage of proficient students over time are likewise presenting an unrepresentative view of distributional change. At Level 2, this article demonstrates that neglecting to imagine trajectories stemming from other proficiency cut scores is akin to ignoring a larger set of trends, all correct, none complete, each capable of encouraging dramatically different substantive interpretations.

The Level 3 argument demonstrates that PPS-based gap trends—changes in gaps that indicate progress toward equity—are susceptible to striking distortions under choices of cut scores. This argument draws heavily on Paul Holland's excellent and undercited exposition (Holland, 2002). Strenuous cautions against using PPS statistics for the reporting of gap trends have been made by others, including Bracey (2006), Koretz and Hamilton (2006), and Linn (2007). Whereas these citations focus on gap trends, I frame this caution as one of many levels of argument against PPS-based statistics.

The Level 4 argument shows that the dramatic dependencies of PPS-based trends, gaps, and gap trends on the selection of cut score are neither fully predictable nor easily remedied. References such as Holland's (2002) article present idealized cut-score dependencies assuming normal test score distributions. I present real data analyses revealing dependencies far surpassing those expected under normal distributions. Cut-score-based trends, gaps, and gap trends can be irreparably unrepresentative in practice, and there is no easy fix in the absence of other data.

Finally, the Level 5 argument is that these dependencies may be used, cynically or nobly, as policy tools. From the first four levels, it follows that the choice of cut score can exaggerate trends, minimize gaps, and, more subtly, focus attention on low-achieving students or muddy comparisons between trends on state tests and NAEP. At Level 5, connections are established between the misleading properties of PPS-based statistics and policy strategies that could, cynically or nobly, take advantage of these properties to achieve a desired policy goal. This hierarchical framework positions educational statistics in interaction with educational policies and demonstrates how limitations born in one sphere can travel to the other.

Large-scale educational statistics make headlines and motivate interventions. Trends, gaps, and gap trends influence perceptions of student achievement, teacher quality, and progress toward equity in educational opportunity. With these high stakes, the selection of a metric that encourages a narrow and unrepresentative perspective on trends, gaps, and gap trends can have far-reaching consequences.

This article builds an argument against the overuse of PPS-based statistics and proposes more robust measures of large-scale progress and gap closure in test scores. The reauthorization of NCLB is approaching, and many states are releasing a new round of accountability results for interpretation. This is a timely opportunity to move past limiting depictions of high-stakes test scores.

## The Blind Men: Limitations of PPS-Based Statistics

The surprising properties of PPS-based statistics are best explained by the fact that there are more examinees near the middle of a test score distribution than there are at the extremes. Figure 2 shows a normal distribution of test scores with the percentage of examinees labeled in evenly spaced categories (divisions of 0.5 standard deviation units). Two hypothetical proficiency cut scores are shown. A Level 1 observation is that the two cut scores report different PPS (93% and 31%, respectively). A Level 2 observation is that, if the distribution shifted uniformly to the right by 0.5 standard deviation units, different proportions of examinees would cross the two cut scores (5% and 19%, respectively). One trend appears much larger than the other because PPS-based trend statistics are confounded by the weight of examinees near the cut score. Both of these trend statistics are correct, yet both are missing the big picture: The distribution is simply shifting uniformly to the right.

If the proficiency cut score is near the mode or center of the distribution, PPS-based trends are expected to be larger. If the proficiency cut score is in the tails, PPS-based trends are expected to be smaller. Over longer periods of time, as a distribution shifts across a cut score, results such as those in Figure 1 can arise. When the center of the distribution is near the cut score, trends will be large. After it passes, trends decrease in magnitude. Cut-score-based trends are analogous to blind men in that each can only observe the examinees crossing it. Each cut score reports a different trajectory, and no single trajectory adequately represents the whole.

These dependencies exist in practice. Figure 3 shows PPS-based trends for Grade 8 NAEP Reading data from the academic years ending in 2005 to 2007. Results of NAEP assessments can be reported in terms of three different cut scores: Basic,

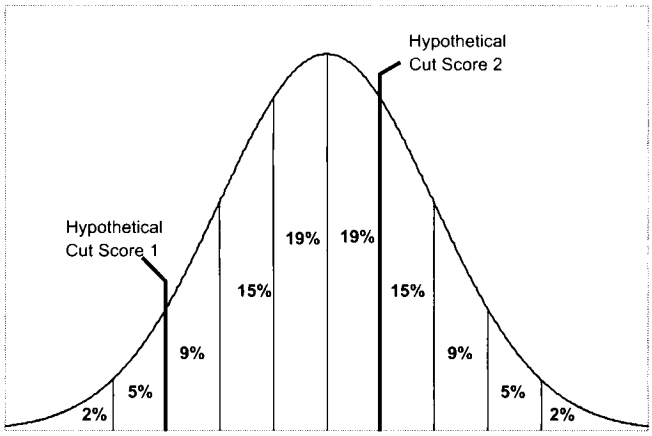


FIGURE 2. *Central cut scores manifest greater trends.*

Proficient, and Advanced. The trends at these three cut scores are shown on a metric of increasing or decreasing percentage points over the 3-year period, and the range of these three trends is shaded. To take one example for illustration, Delaware shows a 3-percentage-point decrease in the proportion of examinees above Basic (b), whereas Delaware's trends for proportions of examinees above Proficient (p) and Advanced (a) are both slightly positive. The width of the bar may also be interpreted as the degree to which the PPS-based trend could change under hypothetical Proficient cut scores between NAEP's Basic and Advanced cut scores.

Figure 3 demonstrates that any reported PPS-based trend is one of many possible PPS-based trends. All of them are correct, but none of them adequately summarizes the big picture. For 36 of these 50 states, different cut-score-based trends will disagree about whether Grade 8 Reading achievement is increasing or decreasing. The widths of bars are similar across the other subject–grade combinations not shown, although NAEP Mathematics scores are generally more positive and less likely to exhibit sign-reversal under a different cut score.

State testing results confirm that the cut-score dependencies of PPS-based trends are not solely a NAEP phenomenon. Figure 4 shows Grade 8 Reading or English Language Arts trends from 16 state assessments from the academic years ending in 2005 to 2007. Data were obtained from state websites. These data also meet a variety of requirements, including reporting percentages above at least three cut scores (or, equivalently, percentages within at least four score categories), unchanged state testing programs, and unchanged cut scores over the 3-year span.

Because states use different terminology to define score categories, the changes in the proportions above the lowest (l), second lowest (m), and third lowest (h) cut score are reported. Figure 4 again demonstrates that trends can vary substantially under selections of cut score, and in many instances, a different selection of cut score can reverse the sign of the trend. This degree of cut-score dependency generalizes across other subjects and grades not shown. The broad range of these statistics reflects the inherent dependency of PPS-based trends on the weight of examinees near the cut score (Figure 2) and begins to hint at the unpredictable ways real-world score distributions can change over time.

Cut-score dependencies are particularly dramatic when they result in sign reversal. Figures 3 and 4 exhibit numerous cases of this sign reversal, which occurs with enough frequency to be coined "trend flipping": a tendency for many PPS-based trends to reverse in sign under selection of alternative cut scores. In contrast, PPS-based gaps rarely flip. Groups that are compared for the purpose of evaluating equity of educational opportunity tend to have differences in average achievement that far surpass the magnitude of a typical trend. Their differences may be summarized by PPS-based gap statistics: the difference between PPS from high- and low-scoring groups. Although PPS-based gap statistics rarely flip, they instead exhibit what may be termed "gap bowing" under selections of alternative cut scores.

As an example, the PPS-based gap between Black and White examinees in 2005 on the National Public NAEP Grade 8 Reading assessment was 30 percentage points at the Basic cut score, 26 percentage points at the Proficient cut score, and 3 percentage points at the Advanced cut score. Just like PPS-based trends, PPS-based gaps are expected to be larger in magnitude near the midpoint between the distributions and smaller at the extremes. A cut score at the midpoint between the modes will maximize the gap by ensuring that the regions with the most examinees for each group are split on either side of the cut score.

In practice, the selection of a state-level cut score involves a carefully considered standard-setting procedure. However, the cut-score dependencies of these statistics are not excused by the appropriateness of a single cut score, even if that cut score could somehow be considered "true." The fundamental problem is that interpretations of PPS-based trends and gaps are confounded with the weight of examinees near the cut score. Even if a cut score could be identified as "true," the trend or gap at that cut score would be a function of the proportion of examinees near it. Any interpretation of a PPS-based trend or gap that fails to account for the weight of examinees near the cut score is assured of being unrepresentative to some degree.

This problem is especially pronounced for gap trends. Gap trends are changes in gaps, requiring four distributions in a differences-of-differences calculation. This topic has been the focus of much of the extant critical literature on PPS-based statistics (e.g., Holland, 2002). Gap trends rise to a Level 3 concern, both because gap trends require four distributions and because the sign reversal of gap trends under different cut scores occurs nearly without exception. The near inevitability of "gap-trend flipping" follows logically from Figure 2: Cut scores near the mode of a distribution will report greater trends. Thus, a cut score near the higher group's mode will be more sensitive to the higher group's progress, resulting in an increasing gap. A cut score near the lower group's mode will be more sensitive to the lower group's progress, resulting in a decreasing gap. As an example, the National Public NAEP Grade 8 Reading Assessment shows an increasing Black–White gap at the Proficient cut score but a decreasing gap at the lower Basic cut score. These reversals are commonplace.

Figure 5 graphs the shape of the dependency of gap trends on cut scores assuming normal distributions shifting equally over time. In this situation, there is no change in the gap between distributions as measured by averages. The left portion of the graph shows that low cut scores will manifest decreasing PPS-based

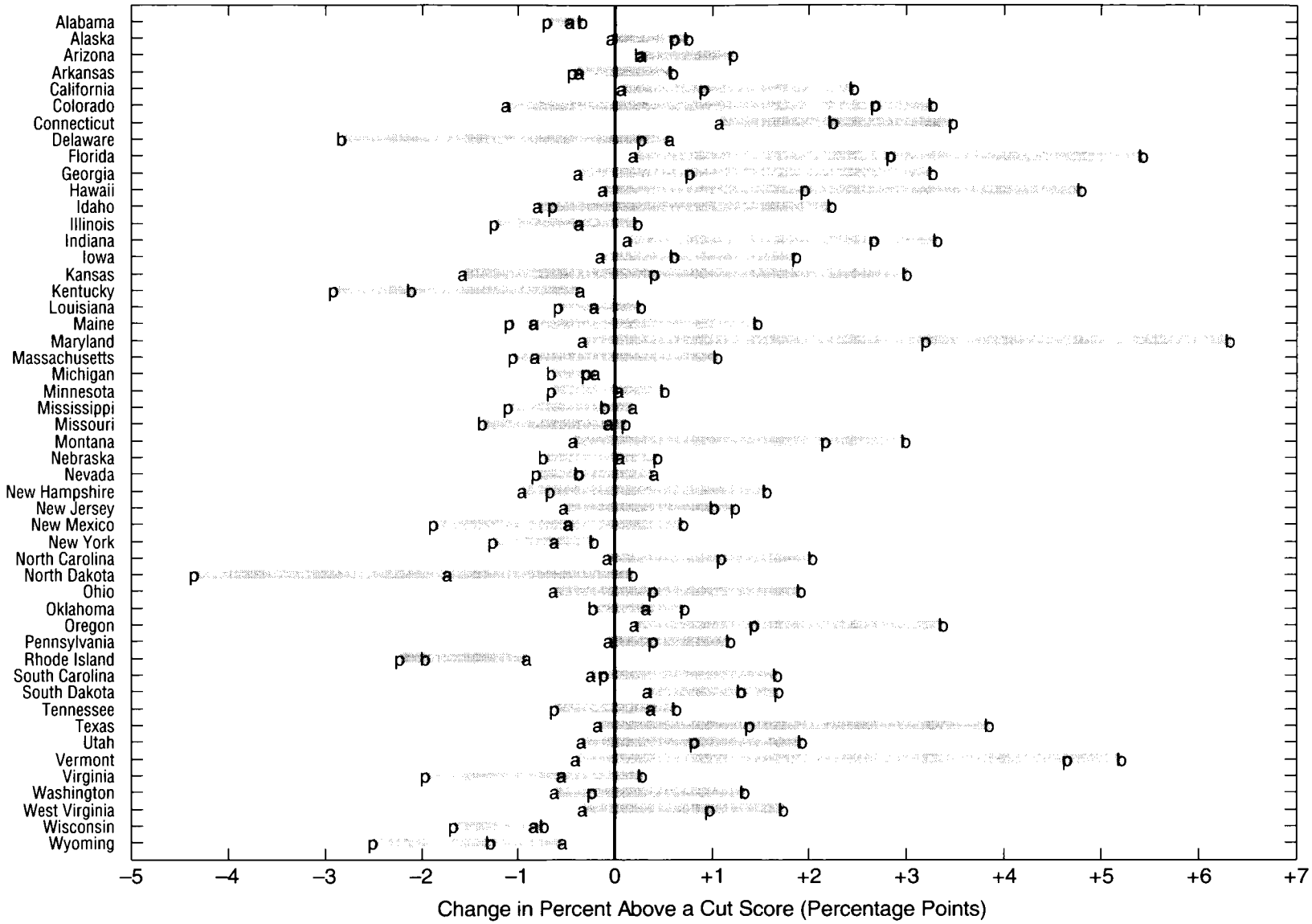


FIGURE 3. *Cut-score dependency of National Assessment of Educational Progress Grade 8 Reading trends, 2005–2007. There are three different cut scores: b = Basic, p = Proficient, and a = Advanced.*

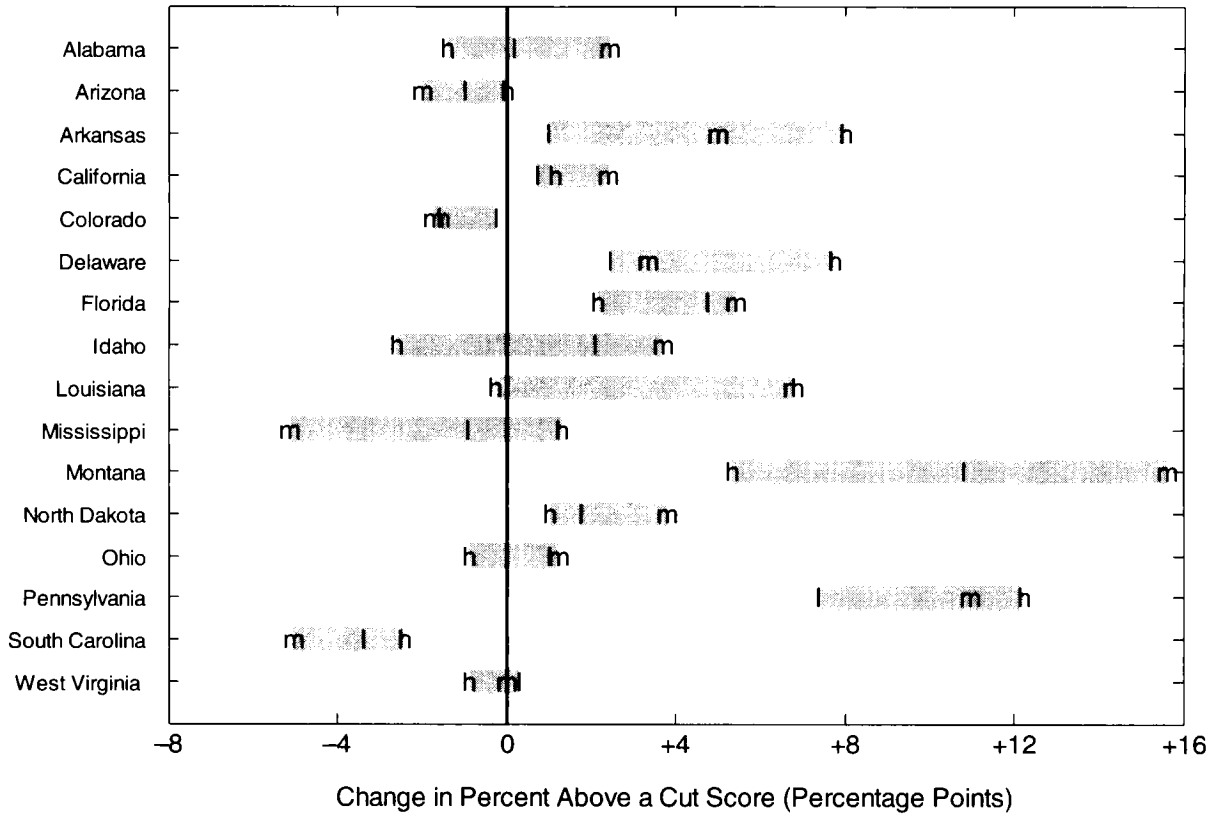


FIGURE 4. *Cut-score dependency of Grade 8 Reading or English Language Arts trends from state tests, 2005–2007. The change in the proportions above the lowest (l), second lowest (m), and third lowest (h) cut score is shown.*

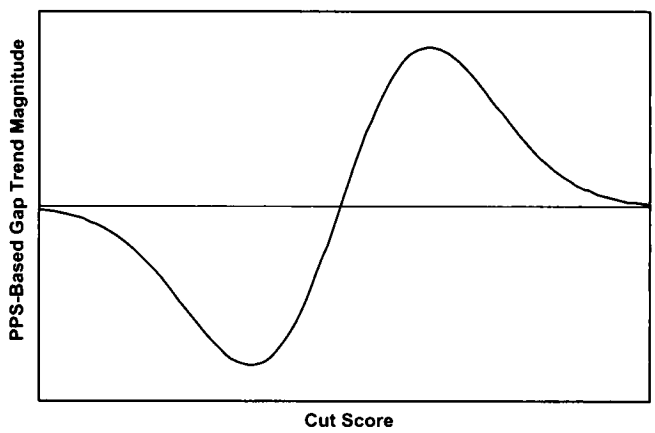


FIGURE 5. *The dependency of Percentage of Proficient Students (PPS)–based gap trends on cut-score choice assuming equal average progress by both groups.*

gaps. The PPS-based gap shows the greatest decreases when the cut score is between the modes of the low-scoring group as it shifts. For higher cut scores, the PPS-based gap will seem to be increasing; it would seem to be largest if a cut score is between the modes of the high-scoring group as it shifts. At either of these extremes, the gap trend is best explained as a cut score revealing greater trends for respective groups simply because the weight of the respective group is near the cut score.

### Level 4: The Irregular Nature of Distributional Change

The previous section has demonstrated how PPS-based trend, gap, and gap trend statistics confound distributional shifts with the weight of examinees near the cut score. These statistics are not incorrect, but a "trend" of +5 percentage points cannot be appropriately interpreted as a trend unless information about the weight of examinees is incorporated. Assuming normal distributions, a trend of +5 percentage points from 90% to 95% would be three times the trend from 50% to 55%. A gap increase of 5 percentage points at one cut score could be a gap decrease of 5 percentage points at another cut score. If normal distributions can be assumed, there is the possibility of a quick fix.

It is possible to infer the weight of students near the cut score based on the proportion above that cut score. If the proportion is near 50%, there are probably many students nearby, for that is the center of the distribution. If the proportion is near 0% or 100%, there are probably very few students nearby. If many students are likely to be nearby, a large trend is expected and should be adjusted downward, and vice versa. If these assumptions about the number of nearby students were accurate, the adjustments would be accurate as well.

These adjustments can be performed by assuming that the distributions of interest are normally distributed with equal variances. The appropriate approach under this assumption involves taking the inverse normal transformation of each of the PPS statistics and then interpreting the differences as effect sizes or, in

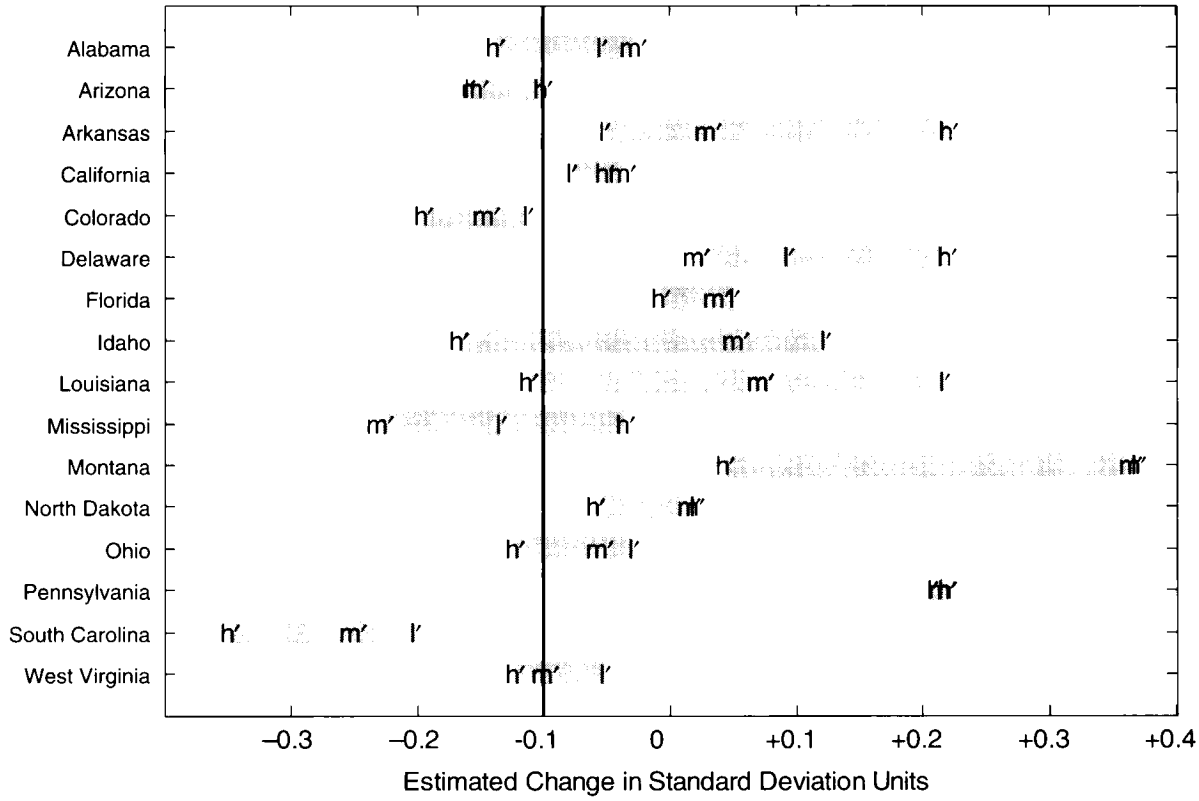


FIGURE 6. *Cut-score dependency of effect size estimates transformed from the Percentage of Proficient Students–based trends in Figure 4. In this figure, the transformed changes above the lowest (l'), second lowest (m'), and third lowest (h') cut scores are shown.*

the case of gap trends, interpreting the difference-of-differences as the change in effect sizes. A simple trend of 50% to 60% proficient can be adjusted in Microsoft Excel by entering the following formula:

$$= normsinv\,(0.6) - normsinv\,(0.5). \quad (1)$$

This returns a value of around 0.25. This is interpretable as an effect size or a normal distribution that shifted one quarter of a standard deviation unit. If the normal assumption held, a data analyst using this transformation would find a trend of 0.25 for all possible cut scores. For example, if the proportion above a higher, fixed cut score increased from 40% to 50%, Equation 1 would also return 0.25 as long as the normal assumption held. Under this assumption, all PPS-based statistics would adjust nicely to a constant, and cut-score dependencies would thereby be remedied. Through transformations to an effect size metric, users of PPS-based trend, gap, and gap trend statistics may hope to disentangle trend, gap, and gap trend interpretation from the weight of students near the cut score. In practice, however, this hope is largely false, and the degree to which it is false is a testament to the irregular nature of distributional change in educational testing.

The results in Figure 6 are calculated from the same state trends as Figure 4, except that all have been transformed by Equation 1. If state distributions followed the normal-shift, equal-variance model, trend estimates within each state would be the same, regardless of the cut score, and the widths of all the bars would be zero. Instead, it is clear that the cut-score dependency of trends persists. Adjusting for the expected weight of students near cut scores by assuming normal distributions will generally not prevent dramatic cut-score dependencies in empirical data. Application of Equation 1 to the NAEP trends in Figure 3 yields similarly cut-score-dependent results.

The shortcomings of the inverse-normal transformation stem from the unrealistic application of the normal, equal-variance model. The ostensible purpose of NCLB is to advance equity and thus reduce the variability of score distributions over time. Further, accountability pressures are not uniform across the distribution, and they are arguably greatest just below the proficiency cut score of state tests. I return to this point later, but for now it suffices to note that normal distributions shifting without changing shape is unrealistic in the NCLB era, and PPS-based statistics cannot be salvaged as a result. Better depictions of trends, gaps, and gap trends are necessary for reporting school, state, and national progress under NCLB.

## The Elephant: Distribution-Wide Perspectives on Distributional Change

Reducing the complexities of distributional change to a single number always involves sacrificing information. As I have demonstrated, PPS-based statistics also sacrifice perspective, and inaccurate inferences are likely to result. If a single summary statistic is to be chosen as a foundation for reporting, it should be the average. Averages incorporate the value of every score into their calculation, and their statistical properties allow scale-up to

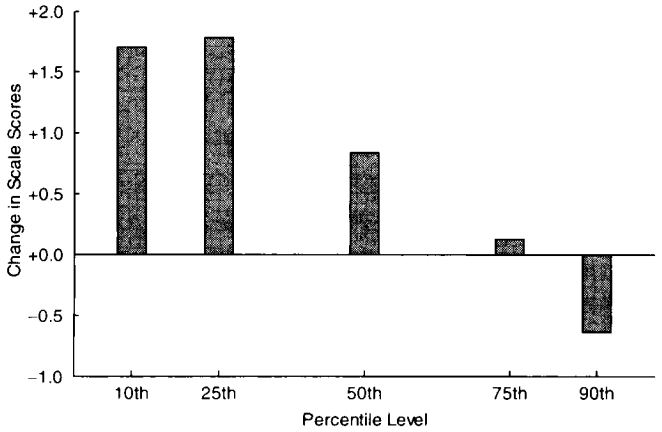


FIGURE 7. *The percentile-based trend by percentile level for the National Public Grade 8 Reading National Assessment of Educational Progress, 2005–2007.*

more advanced analyses. Trends, gaps, and gap trends can be expressed as effect sizes: the difference of averages in standard deviation units. As abstruse as standard deviation units may be to the lay user, the cost of marketing the effect size metric is minimal compared with the cost of the confusion and misinformation induced by a metric that allows trends and gap trends to reverse in sign.

The use of PPS has so dominated NCLB reporting that means and standard deviations are seldom reported by state testing programs. A Center on Education Policy analysis of post-NCLB trends noted that fewer than one half of the states had appropriate data readily available for state-level effect size analyses (Center on Education Policy, 2007), and cross-state analyses of trends and gap trends under NCLB have had to rely on PPS-based statistics to obtain the semblance of a common metric (e.g., Fuller, Wright, Gesicki, & Kang, 2007; Lee, 2006). The first portion of this article demonstrates that this common metric is in fact an illusion. PPS-based trend statistics have little basis for commonality with themselves under different cut scores, let alone with other states. The widespread reporting of PPS-based statistics at the expense of averages impedes not only public understanding of trends and gap trends but also serious education research endeavors.

It is certainly reasonable to ask questions about trends at different locations of the distribution. However, PPS-based trends do not address this question appropriately. Changes in percentages at high and low cut scores do not solely reflect trends for high- and low-achieving students but are instead confounded with the weight of examinees near these cut scores. In contrast, changes in percentiles are not confounded. Figure 7 shows trends at different percentiles for the National Public trend for NAEP Grade 8 Reading. These trends are reported in terms of changes in NAEP scale scores, and they show that gains are more positive for students in the bottom half of the distribution.

If the distribution does not change shape as it shifts, this plot is flat; that is, the trend does not depend on the percentile level. Contrast this with the cut-score dependencies shown in Figure 5, where the statistic exhibits dependencies even under normal assumptions. Holland (2002) argues convincingly that these properties support the use of percentile- rather than PPS-based graphical reporting. Similar distribution-wide plots and perspectives have been proposed by Braun (1988), Ho and Haertel (2006), Livingston (2006), McLaughlin and Bandeira de Mello (2005), and Spencer (1983). State reporting of various percentiles, for example, the 10th, 25th, 50th, 75th, and 90th percentiles that are reported by NAEP, would enable researchers to make significant strides toward a distribution-wide perspective on large-scale educational progress.

### Level 5: Proficiency as a Policy Tool

The PPS acts as the cornerstone statistic of NCLB both by setting a goal (100% proficiency by 2014) and by forming the basis for decisions about which schools will face sanctions. If a school's PPS or any of its subgroups' PPS does not meet the Annual Measurable Objective (a state-specific percentage that increases to 100% by 2014), the school is identified as in need of improvement. Although safe-harbor provisions and, for some states, the Growth Model Pilot Program (U.S. Department of Education, 2005) change this landscape slightly, the core accountability mechanism of NCLB remains one of motivating schools to bring students to proficiency.

The location of the proficiency cut score can function, by accident or by design, as a policy tool. At Level 5, the selection of the proficiency cut score is explicitly framed as a policy decision with substantive consequences. The motivations for cut-score selection can interact with the statistical dependencies presented previously. At Level 1, for example, a low cut score could be selected to result in a large reported PPS. At Levels 2 and 3, a cut score could be selected to maximize trends, minimize gaps, or maximize gap reduction, at least by appearances and at least in the short term. (Interestingly, no single cut score could satisfy all of these priorities at once.) These motivations are certainly cynical; no standard-setting process explicitly includes these considerations. However, regardless of the soundness of the standard-setting procedure, the cut score it generates may maximize, minimize, or distort trend and gap interpretations.

The selection of a cut score can also be motivated by its predicted impact on classroom practice. A recent accumulation of quantitative and qualitative evidence supports the hypothesis that "bubble kids"—students just below the proficiency cut score—receive disproportionate classroom attention and make larger score gains under NCLB (Booher-Jennings, 2005; Choi, Seltzer, Herman, & Yamashiro, 2007; Neal & Schanzenbach, 2007). There are particular concerns if school and teacher resources are "zero-sum," that is, if increases in gains for bubble kids must coincide with decreases in gains for other students and sum to zero. Schools with limited resources may face the greatest pressure to focus on bubble kids in a zero-sum manner, simply because there are no surpluses available for broader allocation. A high cut score may therefore serve to increase real score disparities, as the lowest scoring students face triage.

Under this zero-sum model, responses to NCLB would not be consistent with Title I, improving the academic achievement of the disadvantaged, and would be more accurately described as improving the academic achievement of the near-proficient.

Realizing this, one could set a lower cut score deliberately, not to increase Level 1 PPS statistics (although it would) but to focus resources on the lowest achieving students and the proficiencies developed at that region of the score scale.

Although this may be a noble motivation, the interpretations of PPS-based statistics under this zero-sum model would still be distorted. Generalizations from the proficiency-based trend to full distributional change would be inappropriate, whether adjusted or not. This example crosses all five levels of the hierarchical argument of this article. A cut score may be used as a policy tool to focus attention on students in a particular score region (Level 5), resulting in a higher or lower PPS (Level 1). School responses to bubble kids may disproportionately augment trends at proficiency (Level 4), and this will lead to distorted inferences about trends and gap trends (Levels 2 and 3).

If cut-score-based policies are employed, their success and failure cannot be evaluated by change at that cut score alone. The old carpenter's adage "Measure Twice, Cut Once" can be altered in this context to "Cut Once, Measure Everywhere." Distribution-wide representations like Figure 7 could help to demonstrate that educational progress is not a zero-sum game across students, or they may help stakeholders decide that school responses targeting bubble kids are reasonable concessions as long as trends are generally positive. PPS-based analyses are at best inadequate and at worst inappropriate for addressing these significant research priorities.

### *Proficiency and Cross-Test Comparisons*

The NCLB theory of action does not concede the zero-sum possibility. It considers a cut score not as a lens to focus attention on a particular bubble but as a carrot that motivates everyone below it in the name of rigor and high standards (Rothstein et al., 2006). Within this framework, low proficiency cut scores undermine the accountability model, and upward pressure on cut-score selection seems appropriate. One form of upward pressure on state proficiency cut scores comes from NAEP comparisons. The national assessment has, with few exceptions and for some time, reported lower percentages of proficient students than states for matched subjects and grades, implying higher cut scores and thereby higher standards (Braun & Qian, 2007; Linn, Baker, & Betebenner, 2002; McLaughlin & Bandeira de Mello, 2005; Musick, 1996). As a cross-test, PPS-based comparisons begin to be used as tools of validation and verification, they also function as policy tools, placing pressure on states to match NAEP standards and NAEP trends.

Cross-test comparisons add an additional dimension of complexity to PPS-based statistics because there are two sets of distributions that are each being summarized in a very limited fashion. At Level 1, cut scores must be set for two different tests, with different sampling designs, content frameworks, policy functions, and audiences. At Levels 2 and 3, there are trends, gaps, and gap trends that are inconsistent and misleading in and of themselves, let alone in comparisons across tests. The Level 4 irregularities that may render PPS-based statistics unrepresentative on one test are not likely to be similar across tests. For the comparison of trends, gaps, and gap trends across tests, graphical procedures should be the starting point, effect sizes should be the metric for comparison, and transformed PPS-based statistics should be a last resort. When more appropriate comparisons still result in discrepancies between state tests and NAEP, a more thorough exploration of the many hypotheses explaining these discrepancies may proceed (Koretz & Hamilton, 2006).

A Level 5 analysis of proficiency observes that the statistical shortcomings of PPS-based statistics do not exist in a vacuum. A policy maker aware of Levels 1 through 4 could use a proficiency cut score to manipulate interpretations of trends and gap trends, focus attention on particular regions of score distribution, or encourage state testing programs to set higher or lower standards. Especially when intentions are good, this may sound like a positive feature of proficiency. Upon closer inspection, however, it becomes clear that the success of these manipulations depends on the misinterpretations that this article has tried to dispel. Interpretations could not be manipulated in this manner if users understood the confounding of PPS-based statistics with cut-score location, if averages and effect sizes were used, or if percentiles were analyzed across the distribution. It is precisely these kinds of percentile-based analyses that are revealing the unintended consequences of NCLB, such as disproportionate gains made by bubble kids (Neal & Schanzenbach, 2007).

## The Future of Proficiency

This article has presented a hierarchical argument against the use of proficiency-based trends, gaps, and gap trends in education research. Although no proficiency-based statistic is technically incorrect, every proficiency-based statistic is short sighted, offering only a piece of the distribution-wide perspective that the analysis of high-stakes test score distributions requires. The limitations are so severe that trend flipping is a frequent possibility, gap bowing is ubiquitous, and the possibility of gap-trend flipping is near guaranteed. These statistical shortcomings have analogs in policy. Just as trends at proficiency are unlikely to extend to trends elsewhere, the effects of policy on near-proficient students are unlikely to extend to effects on other students.

It might be easy to mistake this argument for an argument against standards altogether. I earlier introduced the adage "Cut Once, Measure Everywhere." The act of setting a goal is an important and necessary step for all policies. The act of incentivizing progress only by counting those who achieve the goal is short sighted. As a goal, proficiency should expand the understanding of progress toward proficiency, not limit it. States must define proficiency in sometimes elaborate performance-level descriptors as part of their compliance with NCLB. It is interesting that describing progress is not a component of the law. I have shown that a seemingly obvious choice for describing progress, the change in PPS, is misleading and unrepresentative. There is a stark contrast between the detailed state definitions of proficiency and the PPS-based trends that so insufficiently describe student progress toward and past that proficiency.

To better describe school- and state-level progress, descriptors of average-, percentile-, and effect size–based changes should be developed with the same care as performance-level descriptors have. States could work to address the question: What does a Grade 4, statewide, average improvement of 5 scale score points mean? It is an added convenience that changes in percentiles,

such as those shown in Figure 7, are on the same scale as changes in averages. The advantage of this question is that, unlike PPS-based statistics, the answer does not depend on the selection of the cut score and the weight of students who happen to be near it. The answer could take a form similar to Cohen's (1988) "small, medium, large" effect size guidelines—although these are clearly far too conservative for cross-sectional, within-grade effect sizes. More preferably, the answer should be a description anchored to the domain tested. The technical and substantive advantages of this metric over PPS-based statistics would be substantial. Even for states working within the current PPS-based accountability system, the development and dissemination of a more appropriate scale for school- and state-level change would improve the accuracy of interpretations. A given school's PPS remains a relevant statistic within the NCLB framework, but PPS need not and should not form the basis for trends, gaps, and gap trends.

### *The Growth Model Pilot Program and Proficiency*

At this writing, 11 states have received approval to participate in the Growth Model Pilot Program, an extension of NCLB that allows for the incorporation of individual student test score trajectories into accountability calculations (U.S. Department of Education, 2005). Principles of growth models are also under consideration for the upcoming reauthorization of NCLB (U.S. Department of Education, 2008). Under most of the approved models, a student who makes significant growth toward proficiency counts positively toward a school's accountability calculations even though the student has not yet crossed the proficiency cut score. A full review of these models is beyond the scope of this article, but two points are immediately relevant to the topic of proficiency.

First, these growth models attempt to address a fundamental problem with the current, proficiency-based accountability model: the lack of incentives and recognition for low-status, high-growth students and schools. Growth models remedy disproportionate attention to bubble kids by effectively locating all nonproficient students on the bubble and providing even the least proficient students with potentially attainable goals in the near term. Growth models seem to rectify the primary concern of this article: the overemphasis of a particular region of a distribution that distorts perception of the whole.

Second, unfortunately, growth models represent only a small step toward addressing this concern. Models in the pilot program must work largely within the existing principles of NCLB, including the goal of ensuring that all students are proficient by 2014. It follows that growth standards for nonproficient students will be very high, as "one year of progress for one year of instruction will not suffice" (U.S. Department of Education, 2006, p. 5). Although state growth models differ in their details, all share the logical principle that students just below proficiency have a shorter distance to travel than very low-scoring students who have to make large gains to be counted as "on track." The problem of targeting bubble students is thus alleviated but not remedied by proficiency-based growth models. The high standards on growth may explain why initial implementation of growth models made only slight differences in accountability calculations in many early growth model states (Klein, 2007). For states in which growth standards on the lowest scoring students are very high, growth models cannot be expected to reduce significantly targeted teaching to bubble kids.

Another corollary of the proficiency-based growth framework is that growth results become dependent on cut scores. The notion of a trajectory to proficiency clearly depends on the definition of proficiency. In fact, for most states, if the percentage of proficient students is large, the percentage of students eligible for growth calculations will be small. States with high cut scores are likely to have greater proportions of "on track" students simply because more students are not proficient. Simulations under realistic conditions show that the statewide proportion of "on track" students can range from 1% to 20% under plausible choices of cut scores (Ho & Magda, 2008). In cross-state comparisons of "on track" students, misinterpretations may arise as high "growth" is confounded with high cut scores and large numbers of nonproficient students.

Growth models hold immense promise far beyond what I can outline here. In addressing the concerns raised by this article, however, the growth models of the pilot program represent only incremental improvement and pose numerous challenges. The "bubble" may be broadened by growth models but perhaps not as much as expected. The percentage of on-track students is a terrible summary statistic for growth because of its inverse relationship with the percentage of proficient students. Cut-score-invariant alternatives like average gain scores are appealing but require development and dissemination of vertical scales—a unique challenge (Kolen & Brennan, 2004). An interesting area of future development is norm-referencing growth, which may provide attractive interpretations in the classroom as well as at the state level. Recent work has allowed excellent visualizations of norm-referenced growth across the distribution (Betebenner, 2008).

From progress to growth, the common thread linking alternatives to proficiency-based statistics is that they will take work to explain. The education research community can step up in this regard. Means and standard deviations, the staples of statistical analysis, should be reported by state testing programs by default, and changes along a scale score metric can be anchored to straightforward descriptions of progress. Percentile-based trends can provide a distribution-wide perspective for serious analyses. The effect size metric should become a standard for cross-state and cross-test trend, gap, and gap-trend comparison. If the broad bars and curving dependencies in the figures of this article attest to anything, it is that everyone who cares about large-scale educational progress under NCLB is observing it, at best, with blinders on. As a field, it is time we expanded our vision.

## NOTE

I am grateful for support from the Institute of Education Sciences, U.S. Department of Education (Award Number R902B06007) and dissertation fellowships from the Spencer Foundation and Educational Testing Service. I would also like to thank Edward Haertel and Robert Brennan for their guidance; Tracey Magda, Katherine Furgol, and Bradley Thiessen for their research assistance; and Tony Onwuegbuzie, the *Educational Researcher* editorial board and staff, and my anonymous reviewers for their thorough and helpful comments on my drafts.

## REFERENCES


Betebenner, D. W. (2008). *Norm- and criterion-referenced student growth.* National Center for the Improvement of Educational Assessment. Retrieved June 15, 2008, from http://www.nciea.org/publications/normative_criterion_growth_DB08.pdf

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42,* 231–268.

Bracey, G. W. (2006). The 16th Bracey report on the condition of public education. *Phi Delta Kappan, 88,* 151–166.

Braun, H. I. (1988). A new approach to avoiding problems of scale in interpreting trends in mental measurement data. *Journal of Educational Measurement, 25,* 171–191.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313–338). New York: Springer.

Center on Education Policy. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Retrieved September 15, 2007, from http://www.cep-dc.org/index.cfm?fuseaction=Page.viewPage&pageId=495&parentID=481

Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2007). Children left behind in AYP and non-AYP schools: Using student progress and the distribution of student gains to validate AYP. *Educational Measurement: Issues and Practice, 26*(3), 21–32.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher, 36,* 268–278.

Hambleton, R .K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education and Praeger.

Ho, A. D., & Haertel, E. H. (2006). *Metric-free measures of test score trends and gaps with policy-relevant Examples* (CSE Tech. Rep. No. 665). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST). Retrieved April 5, 2007, from http://www.cse.ucla.edu/products/reports/r665.pdf

Ho, A. D., & Magda, T. R. (2008, March). *The dependency of growth models on proficiency standards.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27,* 3–17.

Klein, A. (2007). Impact is slight for early states using "growth." *Education Week, 27*(16), 24–25.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education and Praeger.

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends.* Cambridge, MA: Civil Rights Project. Retrieved September 15, 2007, from http://www.civilrightsproject.ucla.edu/research/esea/nclb_naep_lee.pdf

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved March 15, 2007, from http://epaa.asu.edu/epaa/v11n31/

Linn, R. L. (2007). Validity of inferences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education, 19,* 5–15.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3–16.

Livingston, S. A. (2006). Double P-P plots for comparing differences between two groups. *Journal of Educational and Behavioral Statistics, 31,* 431–435.

McCabe, M. (2006). State of the states. *Education Week, 25*(17), 72–76.

McLaughlin, D., & Bandeira de Mello, V. (2005, June). *How to compare NAEP and state assessment results.* Presented at the 35th Annual National Conference on Large-Scale Assessment. Retrieved April 18, 2007, from http://38.112.57.50/Reports/LSAC_20050618.ppt

Musick, M. (1996). *Setting education standards high enough.* Atlanta, GA: Southern Regional Education Board.

Neal, D., & Schanzenbach, D. W. (2007). *Left behind by design: Proficiency counts and test-based accountability.* University of Chicago. Retrieved July 30, 2007, from http://www.aei.org/docLib/20070716_NealSchanzenbachPaper.pdf

Rothstein, R., Jacobsen, R., & Wilder, T. (2006, November). "Proficiency for all"—An oxymoron. In *Examining America's commitment to closing achievement gaps: NCLB and its alternatives.* Symposium conducted at the meeting of the Campaign for Educational Equity, New York, NY. Retrieved April 18, 2007, from http://www.epinet.org/webfeatures/viewpoints/rothstein_20061114.pdf

Spencer, B. (1983). On interpreting test scores as social indicators: Statistical considerations. *Journal of Educational Measurement, 20,* 317–333.

U.S. Department of Education. (2005, November 15). Secretary Spellings announces growth model pilot, addresses chief state school officers' annual policy forum in Richmond [U.S. Department of Education Press Release]. Retrieved February 12, 2007, from http://www.ed.gov/news/pressreleases/2005/11/11182005.html

U.S. Department of Education. (2006, January 25). *Peer review guidance for the NCLB growth model pilot applications.* Retrieved June 15, 2008, from http://www.ed.gov/policy/elsec/guid/growthmodelguidance.pdf

U.S. Department of Education. (2008, April 22). U.S. Secretary of Education Margaret Spellings announces proposed regulations to strengthen No Child Left Behind. *U.S. Department of Education Press Release.* Retrieved June 15, 2008, from http://www.ed.gov/news/pressreleases/2008/04/04222008.html

Wallis, W., & Steptoe, S. (2007, June 4). How to fix No Child Left Behind. *Newsweek, 169*(23), 34–41.


## AUTHOR


ANDREW DEAN HO is an assistant professor in Psychological and Quantitative Foundations at the University of Iowa, 316 Lindquist Center, Iowa City, IA 52242; *andrew-ho@uiowa.edu.* His research interests are in statistical methods and conceptual frameworks for validating test-score gains under high-stakes accountability systems.



Manuscript received June 6, 2007
Final revisions received June 26, 2008
Accepted July 17, 2008