Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples

CSE Report 665

Andrew D. Ho and Edward H. Haertel
CRESST/Stanford University

January 2006

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

Project 3.5 Methodological Issues in Accountability Systems, Strand 1: The Behavior of Linking Items in Test Equating (1st 2 years).

Project Directors: Andrew D. Ho and Edward H. Haertel, CRESST/Stanford University

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences or the U.S. Department of Education.

.

METRIC-FREE MEASURES OF TEST SCORE TRENDS AND GAPS WITH POLICY-RELEVANT EXAMPLES

Andrew D. Ho and Edward H. Haertel
CRESST/Stanford University

Problems of scale typically arise when comparing test score trends, gaps, and gap trends across different tests. To overcome some of these difficulties, we can express the difference between the observed test performance of two groups with graphs or statistics that are metric-free (i.e., invariant under positive monotonic transformations of the test score scale). In a series of studies broken into three parts, we develop a framework for the application of metric-free methods to routine policy questions. The first part introduces metric-free methodology and demonstrates the advantages of these methods when test score scales do not have defensible interval properties. The second part uses metric-free methods to compare gaps in Hispanic-White achievement in California across four testing programs over a 7-year period. The third part uses metric-free methods to compare trends for "high-stakes" State Reading test scores to State score trends on the National Assessment of Educational Progress from 2002 to 2003. As a whole, this series of studies represents an argument for the usefulness of metric-free methods for quantifying trends and gaps and the superiority of metric-free methods for comparing trends and gaps across tests with different score scales.

Part I: A Metric-Free Framework

As federal educational policies are implemented by states, two questions are being placed as cornerstones of educational decision-making: a) How are test scores changing over time? and b) How are differences between the test scores of certain groups changing over time? To use straightforward terminology, we distinguish here among "change" (Question 1), and "gaps," and "changes in gaps" (Question 2). In smaller, policy-research circles, a third, higher order question arises: Are answers to the first two questions the same for different tests? Studies that address this third question (Klein, Hamilton, McCaffrey & Stecher, 1998; Koretz & Barron, 1998; Linn, Graue, & Sanders, 1990; and many more) have received increased attention as the demand for external validation of high-stakes test score gains has increased.

These questions might be addressed in terms of averages. We can discuss the policy-relevant averages with a running example. Call X a Grade 4 Reading Test score from California's Standardized Testing and Reporting (STAR) system, and call Y a California Grade 4 Reading Test score on the National Assessment of Educational Progress (NAEP)[1]. Subscripts a and b correspond to scores from individuals from an advantaged and a disadvantaged group respectively, and subscripts 1 and 2 correspond to the 2002 and 2003 administrations of the tests respectively.

Score differences are not comparable for tests using different score scales. Effect sizes may be used that "standardize" these differences by expressing them in standard deviation units. When the standard deviations are different for the two groups, we can pool the standard deviations to obtain an estimate of the standard deviation of the common population from which both groups are presumed to be sampled. Let σ be a pooled standard deviation where the first subscript(s) designate the test, and the subscripts in parentheses designate the standard deviations to be pooled. The design is Grade 4-only and thus cross-sectional with successive cohorts at the same grade level; it is not intended to allow for conclusions about the learning of individual students or groups of students over time.

Change (The change in the average STAR test score for all students from 2002 to 2003.):

$$\Delta\overline{X} = \overline{X_2} - \overline{X_1} \tag{1}$$

Gaps (The difference between advantaged and disadvantaged groups' average STAR test scores in 2002.):

$$G_{X1} = \overline{X}_{a1} - \overline{X}_{b1} \tag{2}$$

Changes in Gaps (The change in the gap between the groups on the STAR test from 2002 to 2003.):

$$\Delta G_X = G_{X2} - G_{X1} = \left(\overline{X}_{a2} - \overline{X}_{b2}\right) - \left(\overline{X}_{a1} - \overline{X}_{b1}\right) \tag{3}$$

Trend Discrepancy (The discrepancy between changes in test scores on the STAR test and NAEP.):

$$\frac{\Delta\overline{Y}}{\sigma_{Y(12)}} - \frac{\Delta\overline{X}}{\sigma_{X(12)}} \tag{4}$$

[1] NAEP uses a matrix sampling scheme that does not allow meaningful interpretations of individual scores. Instead, "plausible values" are drawn from an empirically defined distribution of examinee proficiency. To be precise, "individual" NAEP scores Y should be thought of as these "plausible values."

Gap Trend Discrepancy (The discrepancy between the changes in gaps on the STAR test and NAEP.):

$$\left(\frac{G_{Y2}}{\sigma_{Y2(ab)}}-\frac{G_{Y1}}{\sigma_{Y1(ab)}}\right)-\left(\frac{G_{X2}}{\sigma_{X2(ab)}}-\frac{G_{X1}}{\sigma_{X1(ab)}}\right) \quad (5)$$

All of the calculations presented above are dependent on the measurement scale, a fact that has largely been ignored by consumers of these statistics. Spencer (1983) presents an illustrative worst-case scenario of the consequences of ignoring scale when comparing averages which is based on an observation by Lehmann (1955). Spencer presents two groups of five test scores ({10, 10, 20, 45, 50} and {10, 20, 30, 30, 40}) with average test scores of 27 and 26 respectively. He then gives an example of a positive monotonic transformation, $20\_\log_{10}(X)$, that not only changes the magnitude of the gap but *reverses its direction*; the new averages are 26.6 and 27.4 respectively.

Such a transformation would seem inappropriate if we were to believe in the interval properties of the original score scale, but the argument for interval properties of scales in educational measurement is often untenable. Two justifications for interval properties, Rasch scaling and scaling that produces a normal distribution, do not withstand rigorous scrutiny (Lord, 1975, 1980; Yen, 1986; Zwick, 1992). Rasch scaling results in a score scale that has interval properties with respect to the logits of the probabilities of a correct response to an item. However, as Zwick argues, this interval property is internal to the scale itself and thus does not have implications for the trait that the scale attempts to measure. Further, as Lord shows, the data cannot tell us which of any number of monotone increasing transformations is preferable. Rasch scaling is mathematically convenient but is not in itself a theory that confers interval properties on its resulting scale. Additionally, it has little connection to most teachers' or policy makers' intuitions.

The rationale for scaling to create a normal score distribution holds that most traits are probably more-or-less normally distributed, so normally distributed scores should have a more-or-less linear relation to the underlying trait. Also, normal score distributions are convenient for traditional statistical analyses. Educators and policy makers might prefer a scale with interval properties defined by some external criterion. The grade-equivalent scale, for example, is often interpreted as if equal units represented expected progress over equal intervals of time. In accountability systems, a sort of equal-interval property may be based on the importance attached

to reaching successive achievement levels defined through a judgmental process. Table 1 presents an example, taken from Kentucky's Commonwealth Accountability Testing System (CATS) (Kentucky Department of Education, 2004, p. 23). The 60-point span for the Apprentice category is larger than the 40-point spans for the Novice and Proficient categories; this reflects an emphasis on helping all children reach proficiency.

Table 1

Kentucky's CATS Test Score Scale with Equal-Interval Properties Defined Through a Judgmental Process

Performance Level	Weight
Novice Nonperformance	0
Novice Medium	13
Novice High	26
Apprentice Low	40
Apprentice Medium	60
Apprentice High	80
Proficient	100
Distinguished	140

1.1 Conceptual Framework

We have argued that average-based statistics are susceptible to problems of scale. In particular, comparing two groups' score distributions on the same test by taking the difference of their means yields a statistic that is plastic and may even change sign under certain positive monotone transformations. While similar statements could be made for all of the equations 1–5, equations 1–3, which we call "simple" average-based statistics, nonetheless reflect commonsense understandings of gaps and changes in gaps. Once we look beyond simple average-based statistics, however, the way that we conceptualize changes, gaps, and changes-in-gaps becomes more consequential. For example, two ways of conceptualizing the change-in-gap statistic are equivalent in a simple, average-based framework. The first way, following equation 3, is to ask, how does the gap between two groups change over time? This implies summarizing pairs of distributions at the same time point to get "gap" statistics, and then seeing how that gap statistic changes. The second is to ask,

how does the first group's change over time compare to that of the second group? This implies summarizing pairs of distributions of the same group to get "change" statistics, and then seeing how those change statistics compare.

$$\begin{aligned}\Delta G_X &= G_{X2} - G_{X1} \\ &= (\overline{X}_{a2} - \overline{X}_{b2}) - (\overline{X}_{a1} - \overline{X}_{b1}) \\ &= (\overline{X}_{a2} - \overline{X}_{a1}) - (\overline{X}_{b2} - \overline{X}_{b1}) \\ &= \Delta\overline{X}_a - \Delta\overline{X}_b\end{aligned} \tag{6}$$

In other words, in this case, the change in gap is equivalent to the difference in changes. In contrast, for both the metric-free statistics that we will introduce and the effect size-based statistics that we use as a metric-dependent comparison, the change in gap is not necessarily equivalent to the difference in changes. As an illustration, in an effect size-based framework, the differences between averages must be divided by pooled standard deviations. In the case of the change-in-gap statistic, there are three ways to pool standard deviations that are only equivalent in the unlikely event that the standard deviations of the four groups are equal. The first pools pairs of standard deviations at the same time point, just as the second line of equation 6 obtains "gap" statistics. The second pools pairs of standard deviations from the same group, analogous to how the third line of equation 6 obtains "change" statistics. The third takes the change-in-gap statistic from equation 6 and divides by the pooled standard deviation of all four distributions.

There are theoretical, conceptual, and practical issues to consider in making this choice. Theoretically, pooling should be done when the standard deviations are all estimates of the same parameter. If we believe that the standard deviations of the test scores of different groups are equal and are not expected to change over time, we can pool all four standard deviations. If there is evidence that different groups' population standard deviations differ, then the standard deviations of different groups should not be pooled. Inasmuch as an ostensible purpose of high-stakes testing policies is to reduce score variance over time while increasing the mean, the standard deviations of different groups should perhaps be pooled at individual item points. This corresponds to the first line of equation 6, a change-in-gap formulation.

On balance, we have concluded that standard deviations for different groups at a given point in time are reasonably similar. We also believe that a change-in-gap is more intuitively appealing than the difference-in-changes. Thus, it is this conceptualization that we adhere to throughout the first two parts of this paper. The

metric-free summary statistics we will present will reflect a gap between two groups at the same time point, and we will show how this gap changes over time. In the metric-free framework, it will be clear that this is not always equivalent to operationalizing our metric-free statistics as change and taking the difference in changes. Practically, the change-in-gap formulation also allows for visuals that track the gap across more than two time points; this is more awkward in a difference-in-changes formulation.

This focus on "gaps" as opposed to "changes" may seem off-target with respect to the provisions of the No Child Left Behind (NCLB) Act of 2001. Under NCLB, states must establish Annual Measurable Objectives (AMOs) that take the form of a minimum required percent proficient. These AMOs begin with "starting points" derived according to a formula in the law, and AMOs must rise to 100% by 2014, although they may be held constant for up to 3 years at a time. Adequate Yearly Progress (AYP) requires that for each of reading and mathematics, the percent proficient for the school as a whole and for each numerically significant subgroup must exceed the AMO. Closing the gap is not an explicit target of the law, but meeting AYP in 2014 necessarily means that any gap between numerically significant subgroups will be 0 as measured by percent proficient. While changes in gaps may not be a central policy focus as a matter of law, they are central to the rhetoric of educational improvement and, indeed, central to the stated purpose of NCLB itself:

> The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education This purpose can be accomplished by closing the achievement gap between high- and low-performing children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers. (Public Law #107-110, 2002, Title I, Sec. 1001(3))

We maintain that change-in-gap statistics are essential to questions about the distribution of educational opportunities and, in particular, to the differential impact of high-stakes testing policies on different groups of students. However, the metric-free framework and statistics we present are just as applicable to "changes" as they are to "gaps" and "changes in gaps," and extensions of this framework to metric-free measures of progress for groups of students are straightforward (see Part III).

1.2 Metric-Free Graphs and Statistics

1.2.1 Metric-Free Graphs for Comparing Two Distributions

The cornerstone of our metric-free framework is the Probability-Probability (PP) plot (Gnanadesikan, 1977). We first define a Cumulative Distribution Function (CDF), F(x), which takes a score x and returns a proportion p representing the percentage of students with a test score less than or equal to x. We can then define a PP plot for distributions a and b that returns a proportion $p_b = F_b\,[F_a^{-1}(p_a)]$ for all p_a. Note that a PP plot treats the two groups in a symmetric fashion; if $p_b = F_b\,[F_a^{-1}(p_a)]$ then $p_a = F_a\,[F_b^{-1}(p_b)]$. Figures 1 and 2 demonstrate PP Plot construction for normal distributions a and b with unit standard deviations and means −0.5244 and +0.5244 respectively. The PP plot in Figure 2 highlights the point (0.3, 0.7), and Figure 1 demonstrates how this point is derived from the two CDFs. $F_a^{-1}(0.3) = 0$; in other words, 0 is the score that 30% of group a is at or below. $F_b\,[F_a^{-1}(0.3)] = 0.7$; in other words, 70% of group b is at or below the 30th percentile of group a. The PP plot can be interpreted as showing the proportions of group b examinees at or below given percentiles of group a.

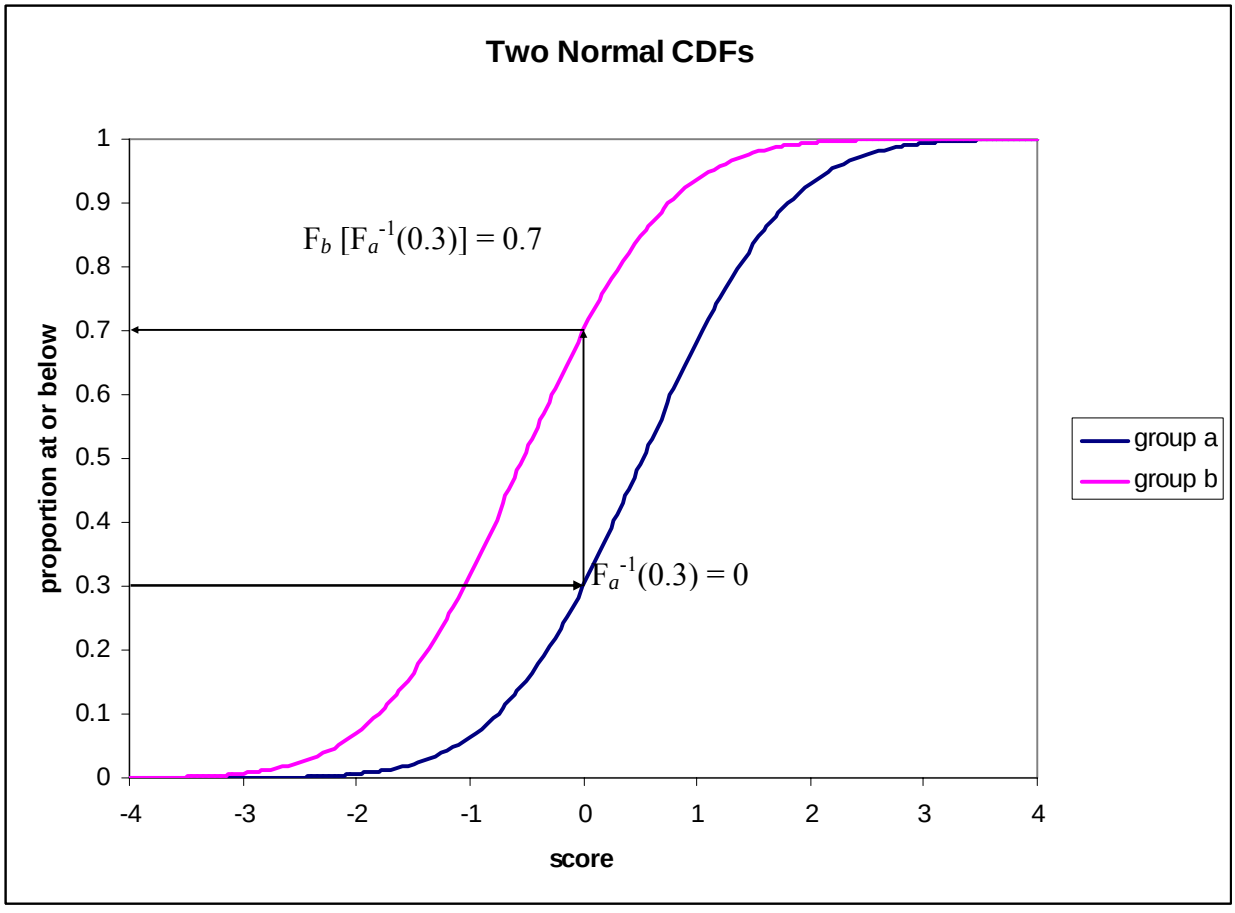


Figure 1. Two Normal CDFs.

Visually, each point on the PP plot can be understood as a point plotted from the two intersections of a vertical "slice" through the two CDFs. No matter how the scale is stretched or transformed horizontally, the intersections of the CDFs with this vertical slice will keep the same values. In this way, PP Plots are metric-free; they are invariant to any positive monotonic transformation of scale. We can also see that PP plots contain a representation of the gap between two distributions. The $p_b = p_a$ diagonal on a PP plot represents two identical CDFs, and PP plots that bulge out from the diagonal represent distributions that are offset. The larger the bulge, the greater the gap between the two distributions. PP Plots have been previously proposed for the purposes of metric-free distributional comparisons by Haertel, Thrash, and Wiley (1978) and by Spencer (1983).

An intuitive visualization of the gap on a PP plot is the orthogonal distance from the curve to the diagonal line. We can use this intuition to define a more visually appealing metric-free plot. For any point (p_a, p_b) on the PP plot, draw a line with slope -1 to the line $p_b = p_a$. The length of this line is simply $\frac{p_b - p_a}{\sqrt{2}}$. A logical reference point for this value is the x-value (or y-value, they are the same) where this line meets the diagonal, that is, $\frac{p_b + p_a}{2}$. Ridding ourselves of the $\sqrt{2}$ scaling factor, we can define a plot that transforms the (p_a, p_b) pairs of the PP plot to ($\frac{p_b + p_a}{2}$, $p_b - p_a$). This plot, which we call a Proportion Difference (PD) plot, is shown in Figure 3.

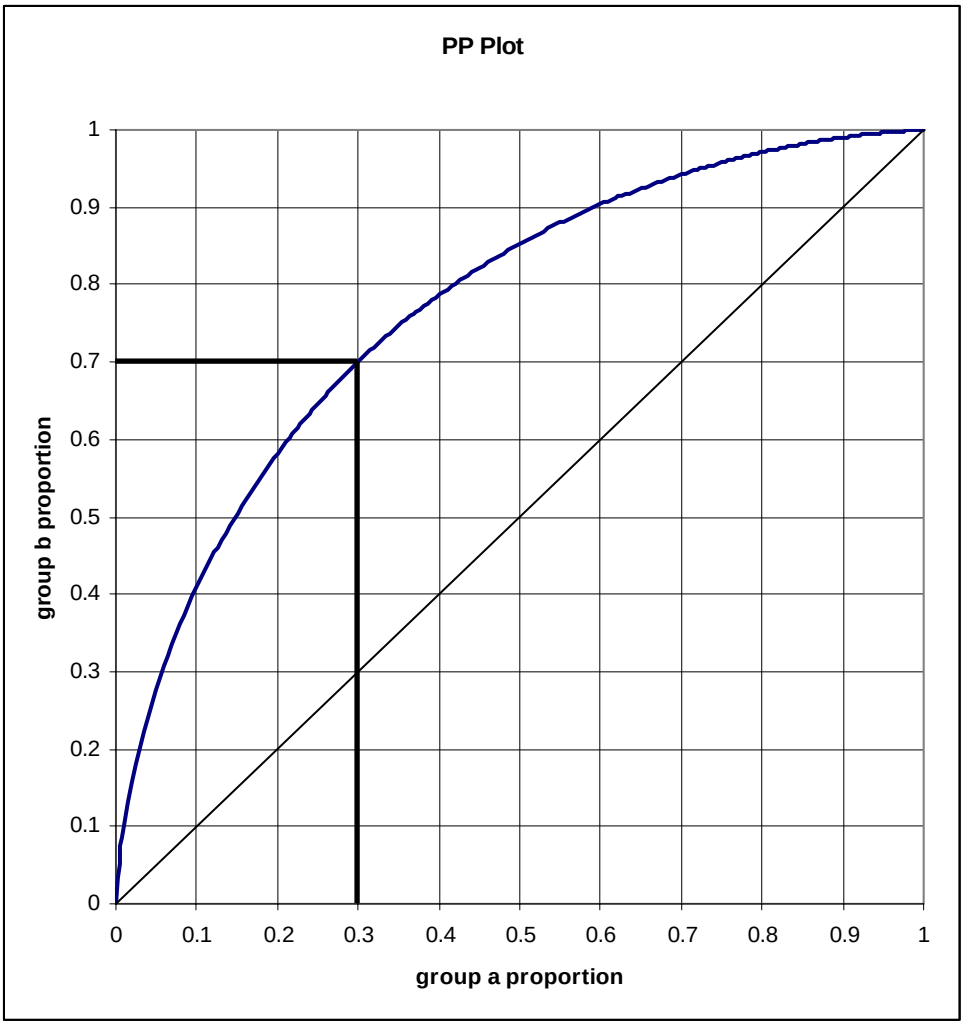


Figure 2. A Probability-Probability (PP) Plot.

A dark line has been drawn to show that the (0.3, 0.7) point from Figure 2 has now become the point (0.5, 0.4). We believe that the PD plot is both visually and substantively more useful than the plot that is generated through the process of "untilting" (Tukey, 1977) which involves subtracting out the diagonal line. If the PP plot in Figure 2 were subjected to untilting, the resulting plot would be skewed.

The PD plot can also be understood as the result of a 45-degree clockwise rotation of the PP plot, followed by a scaling of the x-axis down by $\sqrt{2}$ and a scaling of the y-axis up by $\sqrt{2}$. Beyond its visual appeal and its convenient relationship to the PP plot, the PD plot also affords substantive interpretations. To see this, we observe that the x-values of the PD plot were defined as $\frac{p_b + p_a}{2}$, the "vertical" average of the two CDFs at a given score. Over all scores, we can define a reference

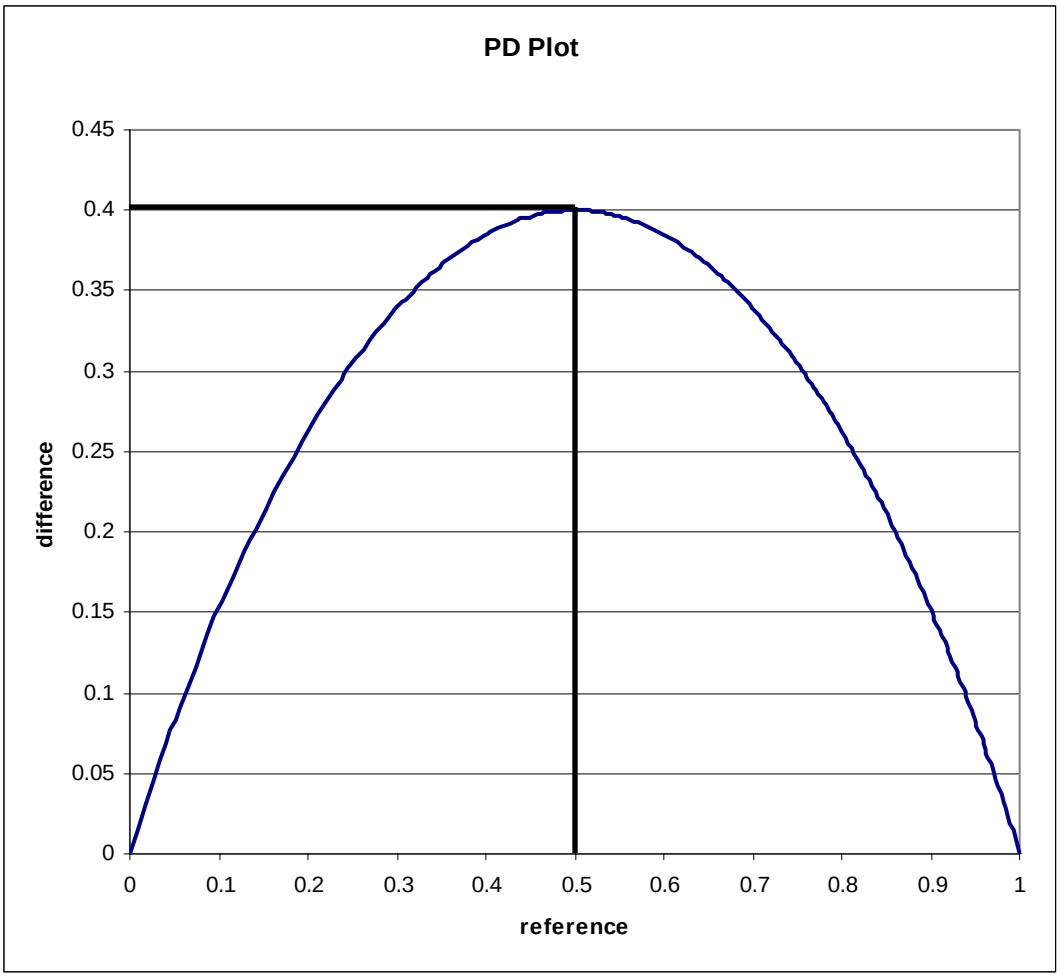


Figure 3. A Proportion Difference (PD) Plot.

distribution *r* as the vertical average of the two CDFs at each score, as shown in Figure 4. The PD plot can be interpreted as the difference between the proportions of group a and group b examinees that are at or below given percentiles of the reference distribution. For example, Figures 3 and 4 show that the difference between the proportions of group a and group b examinees at or below the median of the reference distribution is 0.4. We use the PD plot instead of the PP plot to represent gaps and changes in gaps over time.

1.2.2 Stochastic Ordering

The PP plot in Figure 2 did not cross the $p_b = p_a$ diagonal, and, equivalently, the PD plot in Figure 3 did not cross the x axis. These properties only hold when the two CDFs that define the PP plot are stochastically ordered (Spencer, 1983). Formally,

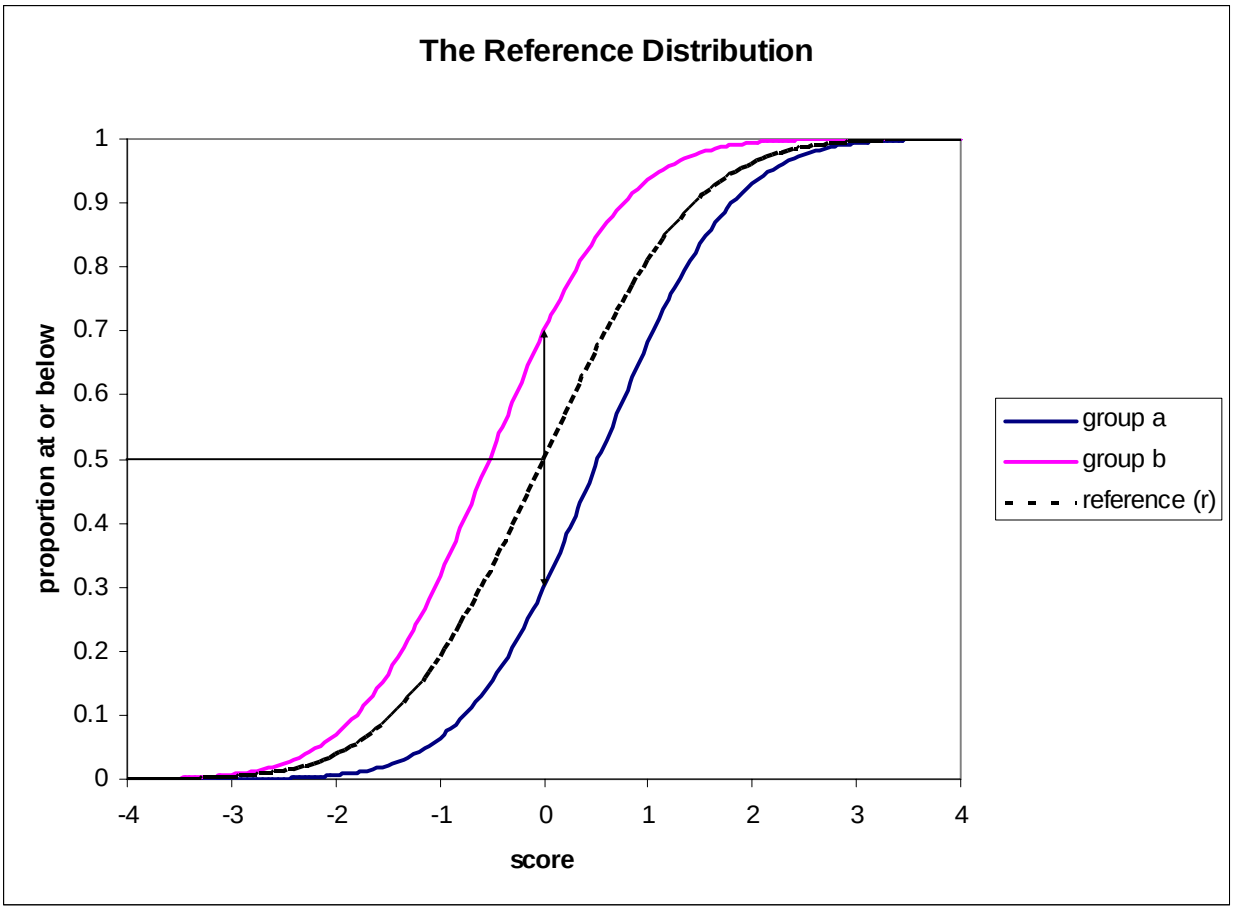


Figure 4. The Reference Distribution.

distribution a is stochastically greater than distribution b when $F_b(x) \geq F_a(x)$ for all scores x. Thus, two distributions are stochastically ordered if and only if their CDFs never cross. Because PP curves that cross the $p_b = p_a$ diagonal are, by definition, indicative of CDFs that cross each other, PP Plots (and PD plots) show at a glance whether distributions are stochastically ordered.

The importance of stochastic ordering is that it acts as a benchmark for the pliability of average-based gap statistics. Average-based rankings of distributions are only invariant to positive monotonic transformations of scale when the distributions are stochastically ordered. In practice, transformation-induced gap reversal is unlikely for most politically significant subgroups on large-scale assessments, even when their distributions are not stochastically ordered. CDFs for advantaged versus disadvantaged groups obtained from NAEP are unlikely to cross at all (i.e., they are likely to be stochastically ordered). If they do cross, it is at a high

or low enough percentile that only an implausibly extreme transformation could reverse the ordering of their averages.

For our purposes, PP plots will always have the advantaged group on the x axis, and PD plots will always report the difference $p_b - p_a$. Because of this and the fact that most of the distributions we deal with are stochastically ordered, both PP plots and PD plots will usually be concave down. It may be worthwhile to note that perfectly normal distributions with different means are stochastically ordered if and only if they have equal standard deviations.

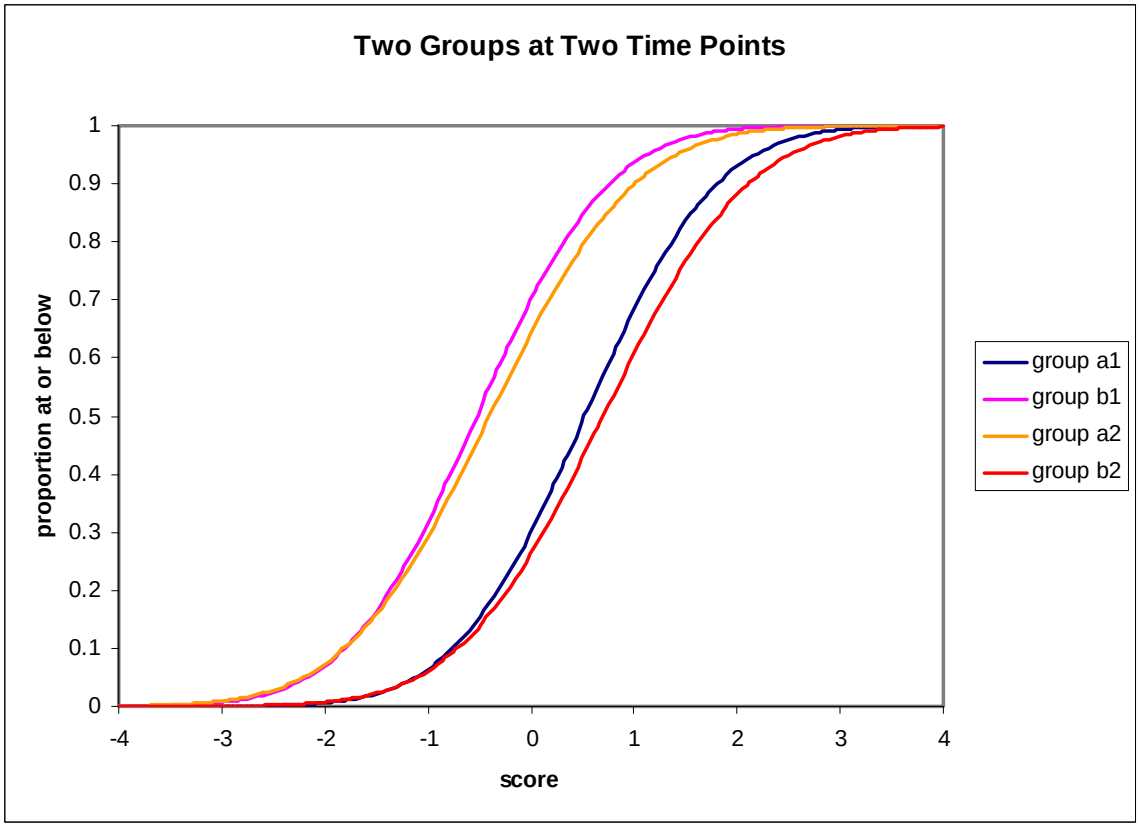
1.2.3 Second-Order Stochastic Ordering

We have noted that if $F_b(x) \geq F_a(x)$ for all scores x, then a is stochastically greater than b, and $\overline{X}_a - \overline{X}_b \geq 0$ under any positive monotone transformation. This notion of stochastic ordering can be generalized from ranking two distributions to ranking the gaps between pairs of distributions. Formally, the gap between the *pair* of distributions *a2* and *b2* is stochastically greater than the gap between the pair of distributions *a1* and *b1* if and only if $F_{b2}(x) - F_{a2}(x) \geq F_{b1}(x) - F_{a1}(x)$ for all scores x. To distinguish the ordering of gaps between pairs of distributions from the ordering of distributions, we will call the former *second-order stochastic ordering*, and we will refer to *gaps* as being stochastically ordered. If the gap at time 2 is stochastically greater than the gap at time 1, then $\left(\overline{X}_{a2} - \overline{X}_{b2}\right) - \left(\overline{X}_{a1} - \overline{X}_{b1}\right) \geq 0$ under any positive monotone transformation.

While the CDFs obtained from large-scale testing programs of advantaged and disadvantaged groups are usually stochastically ordered, the gaps between these groups at different time points are much more likely to be stochastically unordered. By definition, this makes many change-in-gap statistics susceptible to a transformation-induced reversal of sign. This fact has not been well-documented in the literature. It may be interesting to note here that two offset normal distributions with identical standard deviations that both shift in a positive direction will rarely exhibit second-order stochastic ordering. The only exception to this rule occurs when the lower group at time 1 achieves a time 2 mean that is higher than the other group's mean at time 2. In context, this requires the average score of a disadvantaged group to start lower but end higher than the averages of the advantaged groups at both times 1 and 2, something that almost never happens in the context of large-scale testing.

Unfortunately, PP and PD plots do not allow convenient, at-a-glance checks for second-order stochastic ordering. PP plots representing the gaps between groups at times 1 and 2 can be superimposed on each other, but the intuitive check for second-order stochastic ordering would be whether or not these two lines cross. Because each PP plot removes all information about the score scale, the two PP curves cannot "communicate" with each other about common score points to reference. In fact, one advantage to superimposing PP curves is that one PP curve may be derived from a pair of CDFs on one score scale, while the other PP curve derives from a pair of CDFs from a completely different scale. As a corollary, however, the intersections of two PP curves cannot tell us about second-order stochastic ordering of gaps.

Second-order stochastic ordering is best checked by looking at the gaps between pairs of CDFs on the original score scale. The top half of Figure 5 adds two CDFs to Figure 1 that represent the two groups at a second time point. Group *b* has shifted from $N(-0.5244, 1)$ to $N(-.4, 1.1)$ and group *a* has shifted from $N(0.5244, 1)$ to $N(0.7, 1.1)$. Because the standard deviations of the two groups are the same at time 1 and also at time 2, the CDFs of the two groups are stochastically ordered at time 1 and at time 2. However, the bottom half of Figure 5 shows that the curves representing the vertical difference between the groups cross at particular locations on the score scale; this demonstrates that the two gaps do not exhibit second-order stochastic ordering. The simple, average-based change-in-gap statistic is $(0.7 - (-0.4)) - (.5244 - (-0.5244)) = 0.0516$, but because the gaps are not stochastically ordered, there exists a positive monotone transformation of scale where the change-in-gap statistic is negative.



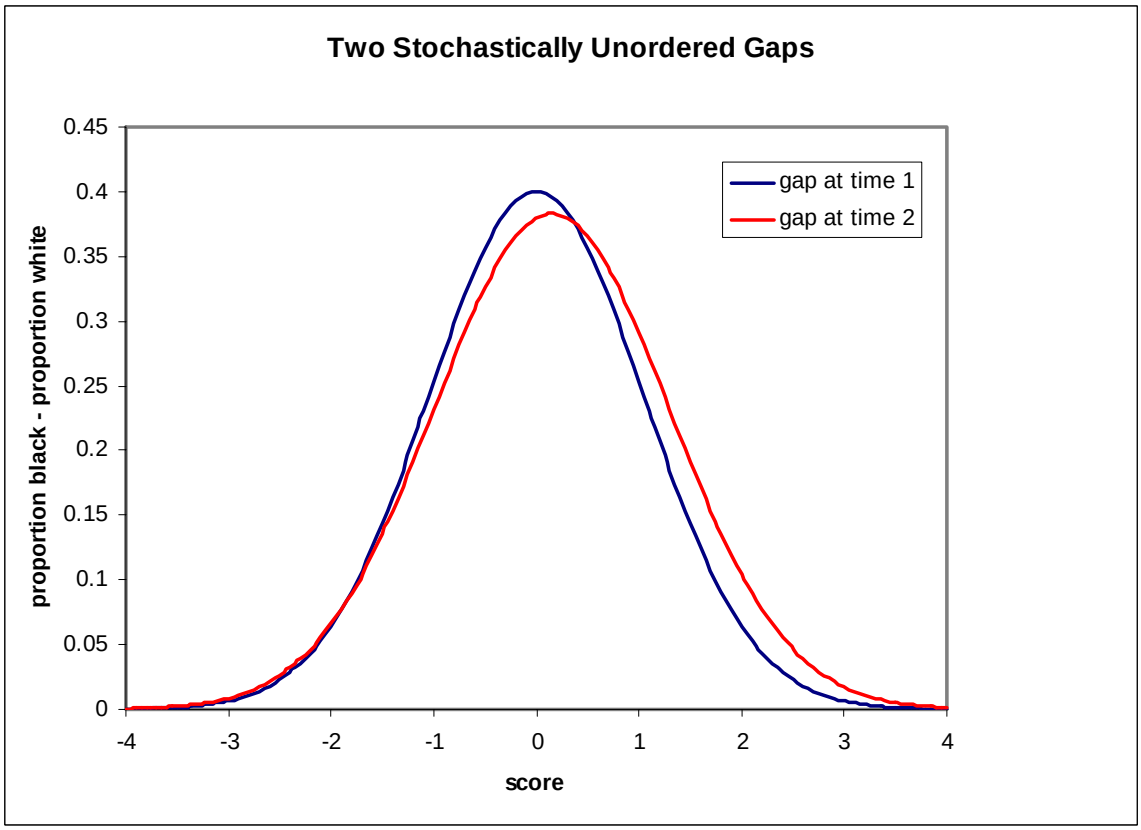


Figure 5. Second-Order Stochastic Ordering.

1.2.4 Internally- vs. Externally-Referenced Metric-Free Gap Statistics

A pervasive alternative to average-based gap statistics is Proportion Above Cut-Point (PAC)-based gap statistics. Cut-points are located on a score scale through one of a number of standard-setting procedures; these cut-points are often labeled in a manner similar to NAEP's "Basic," "Proficient," and "Advanced." A simple measure of a gap, then, is the difference between the proportions of students in two groups who are above the Proficient cutpoint. Formally, this gap can be expressed by $[1 - F_a(x)] - [1 - F_b(x)]$ or, equivalently, $F_b(x) - F_a(x)$ where x is the Proficient cut-point on the score scale. The equivalent PAC-based change-in-gap statistic is simply $[F_{b2}(x) - F_{a2}(x)] - [F_{b1}(x) - F_{a1}(x)]$. We recognize that this corresponds directly to the presentation in Figure 5.

In an important and relevant paper, Holland (2002) demonstrated that PAC-based change-in-gap statistics and median-based statistics (or other percentile-based statistics) can report two seemingly contradictory conclusions about the direction of the change in gaps. In light of our presentation in Figure 5 (analogous to Holland's Figure 7), this is not surprising. Holland describes PAC-based statistics as "vertical" gap measures and percentile-based gap measures as "horizontal" gap measures because of how the gaps are visualized on plots of CDFs. We can see from the bottom half of Figure 5 that the change-in-gap statistic can be dramatically different depending on where the vertical Proficiency cut-point might be. In contrast, because the standard deviations are the same within time points, the horizontal change-in-gap measure is the same as the simple, average-based change-in-gap statistic, 0.0516 over all percentiles. Based on a similar presentation, Holland comes down strongly on the side of using "horizontal," percentile-based statistics.

PAC-based statistics are what we call "externally-referenced" metric-free statistics. The statistic is metric-free because a change of scale cannot influence the vertical distances between the CDFs, but it is externally-referenced because the cut-point is placed for substantive reasons and is the same for all four (or any number of) distributions to be considered. Holland's case and our brief example in Figure 5 are clear evidence that externally-referenced metric-free measures can be misleading. The distinction between our objections is that Holland sees the problem as one where a different externally-referenced cut-point may result in a sign-change, whereas we see that concern as a symptom of the larger issue of second-order stochastic ordering. That is, a positive monotone transformation of scale could

potentially reverse the sign of all traditional change-in-gap statistics, whether they are average-based or percentile-based[2].

In the following section, we present a series of metric-free gap statistics that are "internally referenced," (i.e., the statistics are determined entirely by features of the two CDFs for which a gap is to be measured). These statistics can be derived directly from PP and PD plots that compare pairs of distributions at each time point. We believe these statistics to be superior to both the externally-referenced, PAC-based gap measures Holland finds dubious and percentile- and mean-based measures when there is an essential arbitrariness of scale.

1.3 Metric-Free Summary Statistics

1.3.1 Appropriate Use of Summary Statistics

We propose three metric-free summary gap statistics with caution. Both Holland and Spencer recommend graphical displays for comparing distributions, and we strongly agree. Nonetheless, a scalar summary statistic may also be useful, if only to supplement these types of data displays. In cases where distributions are not stochastically ordered, a metric-free gap statistic may be just as fallible as an average-based calculation in implying that there is a consistent gap between the two CDFs when in fact the CDFs may cross. However, in some senses these metric-free statistics are more up-front; arbitrary scale information has been removed, and a clear explanation of what the statistic represents, especially when supplemented with PP or PD plots, may help to discourage unwarranted conclusions.

1.3.2 The V Statistic

The PP Plot is visually similar to the Lorenz Curve, which is commonly used as a graphical representation of income inequality in economics, and Receiver Operator Characteristic (ROC) Curves, which are now used most often in biomedical fields. The Lorenz curve plots the cumulative percent of income on the cumulative percent of households. Thus, if (0.8, 0.6) is a point on the curve, the bottom 80% of

[2] For completeness, a test for percentile-based notions of second-order stochastic ordering can also be considered. That is, under what conditions is the change-in-gap statistic, as measured by any percentile, positive no matter what the transformation of scale? Any percentile-based gap at time 2 is stochastically greater than any percentile-based gap at time 1 if and only if EITHER $F_{a1}^{-1}(p) - F_{b1}^{-1}(p) \leq 0 < F_{a2}^{-1}(p) - F_{b2}^{-1}(p)$ OR $F_{a1}^{-1}(p) \leq F_{a2}^{-1}(p)$ AND $F_{b2}^{-1}(p) \leq F_{b1}^{-1}(p)$ for any proportion p. Visually, the first condition implies that the horizontal gap at time 2 is positive whereas the horizontal gap at time 1 is negative or zero, and the second condition implies that the two CDFs at time 1 are bounded on both sides by the two CDFs at time 2 (if the first condition does not hold).

households have 60% of the total income. The Lorenz Curve must be convex to the x-axis, because the bottom x% of households cannot have greater than x% of the income or, by definition, they would not be the bottom x%. A commonly derived summary statistic from the Lorenz Curve is the Gini Coefficient. The Gini Coefficient is a statistic between 0 and 1, where 0 represents perfect equality and 1 represents perfect inequality. It can be calculated as the area between the diagonal and the curve divided by the area between the diagonal and the graph edge (0.5 by definition).

We borrow the calculation but not the interpretation of this statistic from its context in economics, as it is clear that the Lorenz curve is conceptually quite distinct from the PP plots we are presenting. We will tentatively call our analog of the Gini Coefficient the *V* coefficient (for de*V*iation from the diagonal). We note that PP curves will cross the diagonal when the distributions are not stochastically ordered, allowing for positive and negative areas. Therefore, a *V* Coefficient of zero may not imply completely overlapping distributions, but that the area on one side of the diagonal may simply cancel out the area on the other side. Further, since PD plots are simply PP plots rotated, then scaled up and down by $\sqrt{2}$, the area under the PD plot is equivalent to the area between the PP plot and the diagonal. Thus, *V* can also be calculated as the area under the PD curve divided by 0.5.

ROC curves are a graphical representation of the tradeoff between false positives and false negatives for a given test, for example, using the level of a protein in a one's blood to predict whether or not one has a particular disease. ROC curves usually plot 1−False Negative Rate against the False Positive Rate. For ROC curves, the Area Under the Curve (AUC) is often used as a summary statistic to approximate the usefulness of a test with respect to minimizing false positives and false negatives. This area corresponds to $\frac{V+1}{2}$. The AUC statistic can also be interpreted as the probability that a randomly chosen diseased patient's level is greater than a randomly chosen healthy patient's level, assuming that greater levels are indicative of disease. This translates to a particularly useful interpretation of the transformed *V* statistic: $\frac{V+1}{2}$ corresponds to the probability that a randomly chosen advantaged student has a higher score than a randomly chosen disadvantaged student.

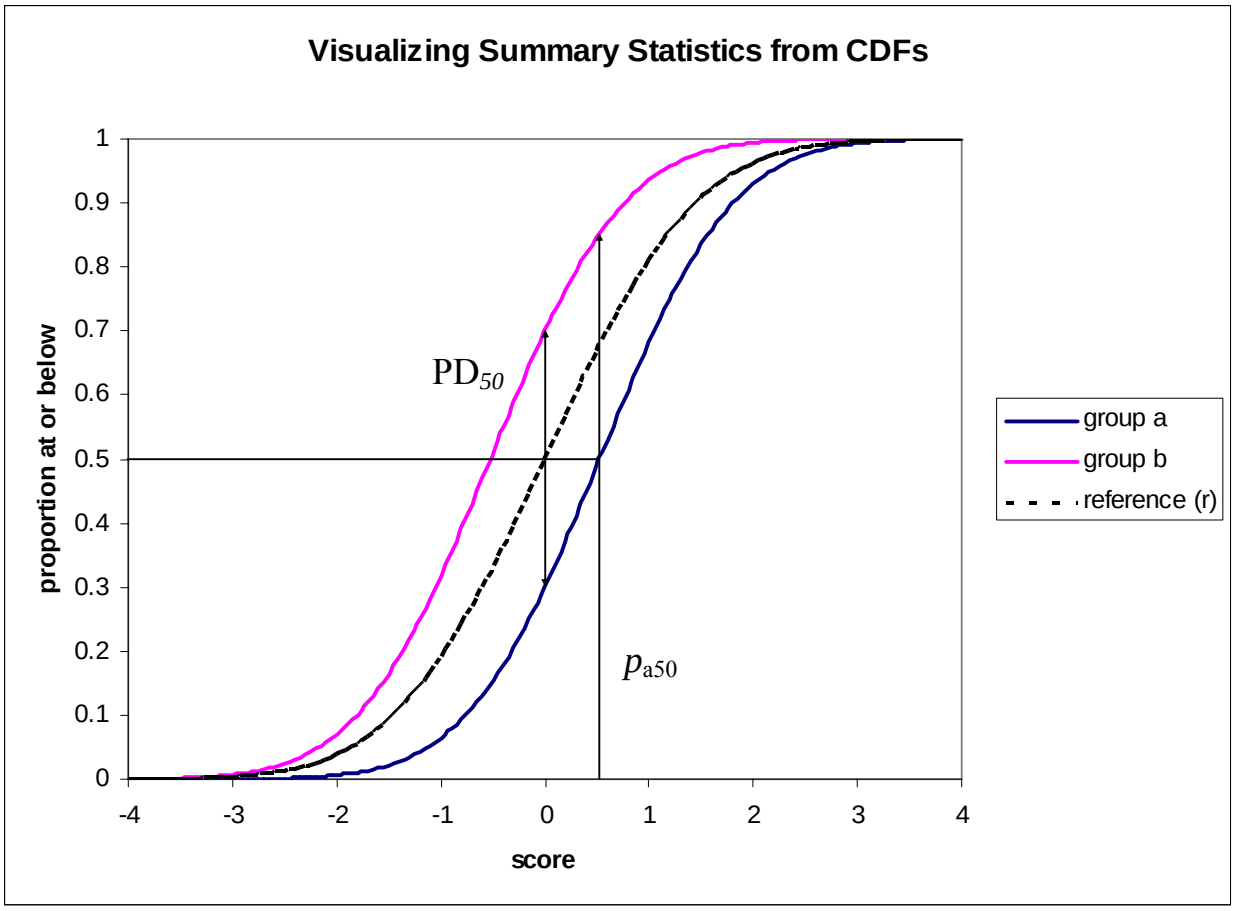


Figure 6. Visualizing Metric-Free Summary Statistics I.

The *V* Statistic is intuitive in the sense that it defines 0 as equality and 1 as inequality, and it is robust in the sense that it takes every point on the curve into account. It may be cumbersome to calculate in practice due to the need for interpolation between what may be sparse data points. The *V* Statistic can be visualized in Figures 7 and 8, though the area shown needs to be divided by 0.5 (or multiplied by 2) to scale *V* appropriately.

1.3.3 The PD50 Statistic

The PD50 Statistic can be defined simply as the height of the PD plot at a reference proportion of 0.5. PD50 is meant to stand for Proportion Difference at a reference proportion of 50%. We have defined the proportions of the reference distribution in a previous section as the average of the two CDFs, $p_r = \frac{p_b + p_a}{2}$ at a given score. Thus, the PD50 Statistic is equivalent to the vertical gap between the two CDFs at the median of this reference distribution. In a PP plot, the PD50 Statistic can

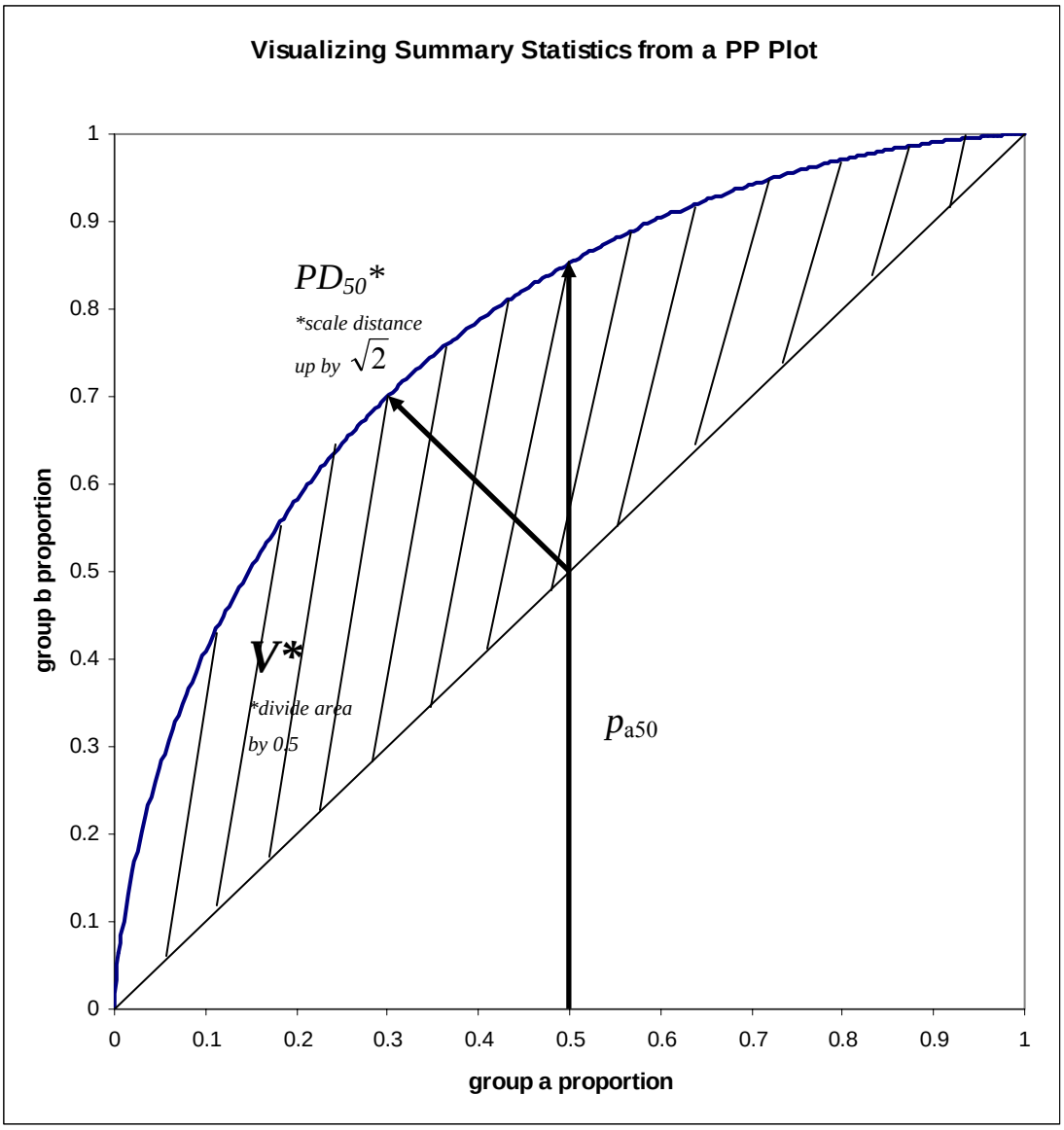


Figure 7. Visualizing Metric-Free Summary Statistics II.

be visualized as the orthogonal distance from the point (0.5, 0.5) to the curve, though this distance needs to be scaled up by $\sqrt{2}$. Figures 6, 7, and 8 all show different visualizations of the PD_{50} Statistic.

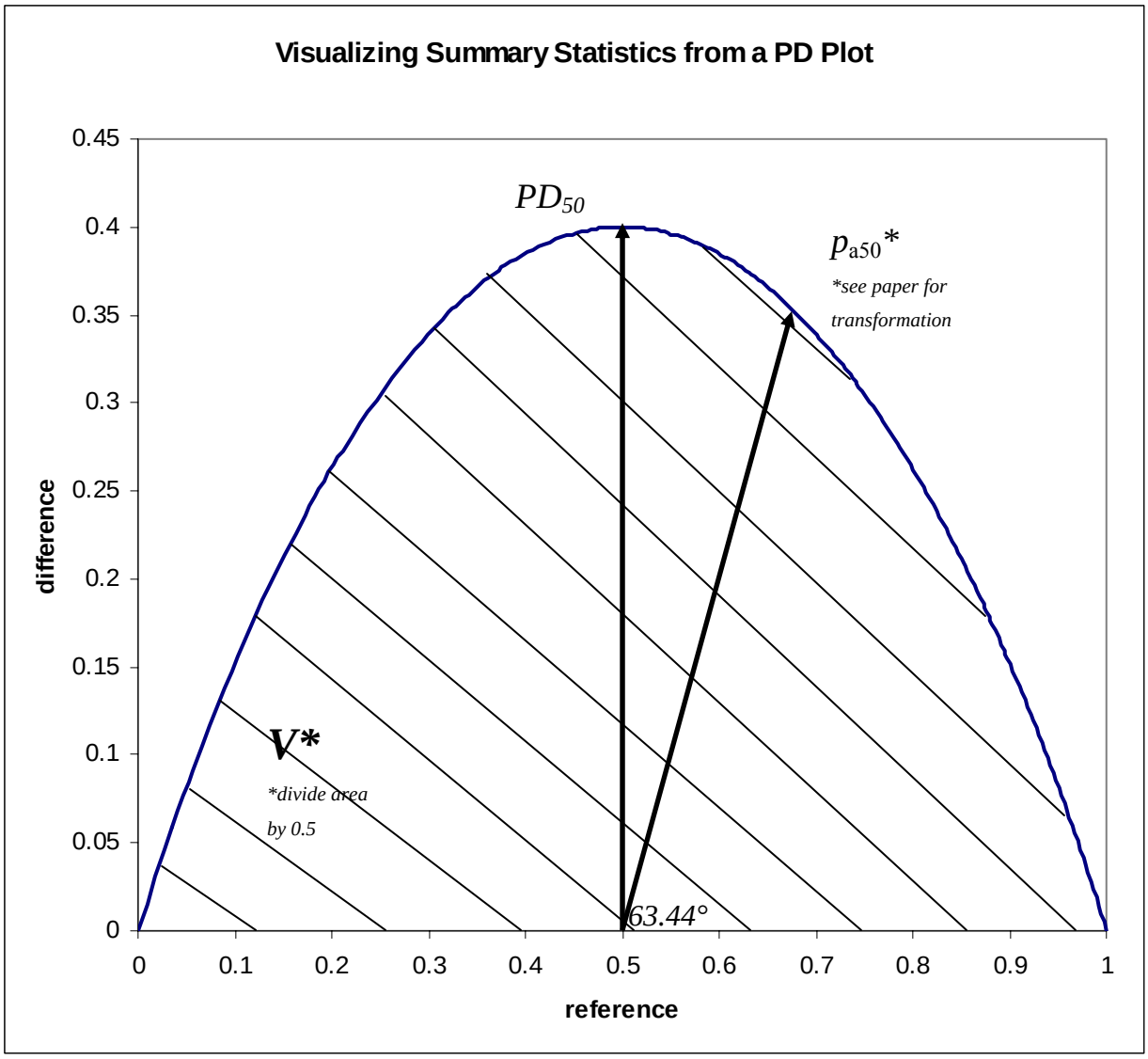


Figure 8. Visualizing Metric-Free Summary Statistics III.

1.3.4 The p_{a50} Statistic

The p_{a50} Statistic is simply the proportion of disadvantaged students at or below the median score of the advantaged students. It is the height of the PP plot at pa = 50%, which gives it its name. The pa50 Statistic can also be visualized on the PD plot, albeit obliquely; it can be found by taking the distance from (0.5, 0) at a 63.435 degree (arctan 2) angle to the curve, and if this distance is b, then pa50 = $\frac{2b}{\sqrt{5}} + 0.5$. This strange transformation simply accounts for the rotation of the PP plot and the subsequent $\sqrt{2}$ scaling factors that result in the PD plot. The p_{a50} Statistic is an

intuitively appealing candidate for a summary statistic due to its ease of interpretation. Figures 6, 7, and 8 all show different visualizations of the p_{a50} Statistic.

1.3.5 Respective Normality

All scale information is lost in the move from two CDFs to a PP or PD plot, but, as we have argued, the arguments for interval scales in educational measurement are usually untenable. It follows that we can pick an arbitrary distribution for one of the groups, and, using a PP or PD plot, uniquely map out what the other distribution must be. By picking a common distribution for all distributions *b*, we can look at the *a* distributions on this common metric. If we choose a standard normal distribution as the common distribution, we can calculate all the statistics from the a distributions that we like. The means of the a distributions could, for example, be interpreted as pseudo-effect sizes. In this way, as we "put back" scale information into metric-free plots, we allow ourselves flexibility of interpretation.

If we reduce PP and PD plots to any of the three summary statistics presented above and set a standard normal b distribution, we do not have enough information to define an a distribution. For example, if we know that the V Statistic is zero and pick a standard normal distribution as the disadvantaged distribution, we still have an infinite number of possible advantaged distributions that can make the areas on both sides of the PP plot diagonal cancel to zero. As a matter of convenience, we can force distribution *b* to be standard normal and force distribution a to be normal with unit variance. Under these assumptions, which we call "respective normality," each of the three summary statistics uniquely defines a mean for distribution *a* which can be interpreted in familiar terms as standard deviation units. We will designate these means, which we will interpret as "standardized" metric-free summary statistics, by adding a ′ (prime) superscript to the names of the statistics from which they were derived.

We label each of these transformed summary statistics as follows:

$$V' = \sqrt{2} * \Phi^{-1}\left(\frac{V+1}{2}\right) \tag{7}$$

$$PD_{50}{}' = 2 * \Phi^{-1}\left(\frac{PD_{50}+1}{2}\right) \tag{8}$$

$$p_{a50}{}' = \Phi^{-1}(p_{a50}) \tag{9}$$

The case of the V' Statistic is particularly interesting. If the two distributions *a* and *b* are normally distributed, the following equation holds:

$$V' = \frac{\overline{X_a} - \overline{X_b}}{\sigma_{(ab)}} \tag{10}$$

Here, $\sigma_{(ab)} = \sqrt{\frac{\sigma_a^2 + \sigma_b^2}{2}}$, the unweighted pooled standard deviation. In other words, if the two distributions are perfectly normally distributed, the V' statistic will be equal to an effect size calculation. This finding has the potential to be misconstrued as circular, as if we have gone from scale-dependent statistics to metric-free statistics and ended up where we started. To the contrary, if the interval properties of a scale are suspect, the effect size calculation can be easily manipulated under positive monotone transformations, whereas the V' statistic will remain constant. Equation 10 shows that if a PP plot happens to have been derived from two perfectly normal distributions (that could just as well be non-normal given an arbitrary re-scaling), the V' statistic will be equal to an effect size calculation on the original distributions. Thus, equation 10 should merely be read as an assertion of a convenient property of the V' statistic: that it can be interpreted, in more ways than the PD_{50}' and the p_{a50}' statistics which reflect only one point on the PP or PD curve, as a metric-free effect size. We acknowledge that even this may be misleading, but until untransformed metric-free gap statistics become familiar in their own right, the "standardized" versions provide a useful bridge to more familiar and established interpretations.

1.4 Metric-Free Summary Statistics in Practice

1.4.1 Comparing Metric-Free and Average-Based Statistics

Comparing metric-free statistics to average-based statistics with scatterplots and correlations is in some sense comparing apples to oranges. We neither expect nor desire a perfect correlation. We have tried to make a case for treating the interval properties of score scales in educational testing with skepticism. If the score scale is suspect, this is a reason to prefer metric-free measures regardless of whether or not they happen to align with average-based gap measures. Further, as the previous section argued, alignment of average-based statistics with metric-free statistics is in no way a defense of the score scale. On the other hand, if there is a rationale laid out for the score scale beyond matters of statistical convenience, average-based statistics may provide more useful substantive information than metric-free statistics.

The following correlations and scatterplots show metric-free statistics in action and demonstrate that the choice of framework does, in fact, make a difference in practice.

1.4.2 Data Sources

Publicly available NAEP data were downloaded with the NAEP Data Tool (National Center for Educational Statistics, 2004). We used data from two grades, Grades 4 and 8, and from six administrations, the 1996, 2000, and 2003 administrations of the State Mathematics Assessment and the 1998, 2002, and 2003 administrations of the State Reading Assessment. At the State level, Black student scores are the most commonly reported scores of all racial or ethnic minorities, so we chose to look at the gaps between White and Black students' score distributions. At each administration, 102 possible gaps could be calculated (for the 50 states and the nation, at Grades 4 and 8). However, some states did not report Black scores due to insufficient numbers, and other states did not participate. Of 612 possible gap statistics (102 x 6 administrations), there was enough information to calculate 430 gaps.

1.4.3 Calculation of Gap and Change-in-Gap Statistics

Average-based gap statistics include a simple difference of means (Equation 2) and an effect size calculated by dividing the difference of means by the unweighted pooled standard deviation (cf. Equation 10). Means and standard deviations were obtained directly using the NAEP Data Tool. Average-based change-in-gap statistics were obtained by simply subtracting time 1 average-based gap statistics from time 2 average-based gap statistics. Thus, a positive change-in-gap statistic indicates an increase in the gap from time 1 to time 2.

Calculation of metric-free gap statistics required numerical methods for interpolation and integration. There are 8 score-proportion data points that can be obtained from the NAEP Data Tool for each distribution. Three are given by the reported proportions above or below the three performance cut-points (Basic, Proficient, and Advanced), and five are given by the five scores corresponding to the 10th, 25th, 50th, 75th, and 90th percentiles. We developed a cubic Bezier-based interpolation function with control points that result in the same curve as the smoothed curve algorithm implemented by graphs in Microsoft Excel. We used this function to obtain Proportion-Proportion pairs for given score points and drew a PP Plot. Extending the plot to the points (0,0) and (1,1), we could use the interpolation

function to calculate the p_{a50}, the PD_{50} and the V statistics. Using the transformations shown in a previous section, we obtained the normalized PD_{50}', p_{a50}', and V' statistics. Metric-free change-in-gap statistics were calculated by subtracting each time 1 metric-free gap statistic from its corresponding time 2 metric-free gap statistic. This was done for every possible time 1 and time 2 pair within Reading and Mathematics, e.g. for Reading, we calculated change-in-gap statistics using 1998 to 2002, 2002, to 2003, and 1998 to 2003 as time 1 and time 2. This resulted in 374 possible change-in-gap calculations.

1.4.4 Comparing Average-Based and Metric-Free Gap Statistics

Table 2 shows pairwise correlations for all five gap statistics. The simple average-based gap statistic shows relatively low correlation with all metric-free statistics and, notably, also with effect size-based gap measures. Effect size statistics show high correlation with metric-free gap statistics, particularly with the V' statistic. This lends further support for the strong conceptual ties between the V' statistic and effect sizes in practice. Of the three metric-free statistics, the p_{a50}' statistic seems least like the others. This may be due to the fact that it samples a non-central location on the PD plot. As the PD point that is sampled deviates from the middle of the plot, it becomes less likely that the distance from (0.5, 0) to this point corresponds to the area beneath the PD curve (see Figure 8). The p_{a50} statistic measures the distance to the PD curve from (0.5, 0) at an angle of approximately 63 degrees. As an extreme example, an angle of 0 degrees will always meet the PD plot at the point (1, 0) and always give an untransformed distance of 0.5 regardless of the actual area under the PD curve.

Table 2

Correlations Between Average-Based and Metric-Free Gap Statistics on NAEP

N=430	Mean Difference	Effect Size	*pa*50′	PD50′
Effect Size	0.8703			
*pa*50′	0.8403	0.9719		
PD50′	0.8756	0.9857	0.9743	
V′	0.8742	0.9946	0.9824	0.9930

1.4.5 Comparing Average-Based and Metric-Free Change-In-Gap Statistics

As we have noted, advantaged and disadvantaged distributions are almost always stochastically ordered in the context of large-scale educational testing. In the previous section, the advantaged group was always stochastically greater than the disadvantaged group, and there was no possible positive monotone transformation that could reverse the ordering of the distributions. In contrast, as we have shown, gaps are often stochastically unordered. As Figure 9 shows, 19.25%, or 72 of the 374 data points, those falling in the second and fourth quadrants, have a change-in-V' statistic that has the opposite sign from the simple, average-based change-in-gap statistic. This is a sufficient but not a necessary condition for a lack of second-order stochastic ordering. In fact, the percentage of stochastically unordered gap pairs is much larger than this result indicates. While some of these unordered gap pairs may only exhibit sign-flipping under implausibly extreme transformations, the 72 change-in-gap statistics where there is this most basic disagreement should be flagged as inconclusive about the direction of the change in gap. As a matter of fact, in a very real sense there is no change in gap to discuss. These are perfect examples of cases where summary statistics, whether average-based or metric-free, are misleading without supplementary graphical displays.

Figure 10 shows a much tighter relationship between changes in effect sizes and changes in V' statistics. This is not surprising given the close relationship between the two statistics shown in Table 2. While the sign miss rate is lower than that shown in Figure 9, none of the 72 cases flagged earlier is redeemed by the findings in Figure 10. Sign discrepancy in Figure 9 is sufficient cause to prompt a closer investigation of the distributions in question, especially if policy decisions or widespread impressions are to be based on these statistics.

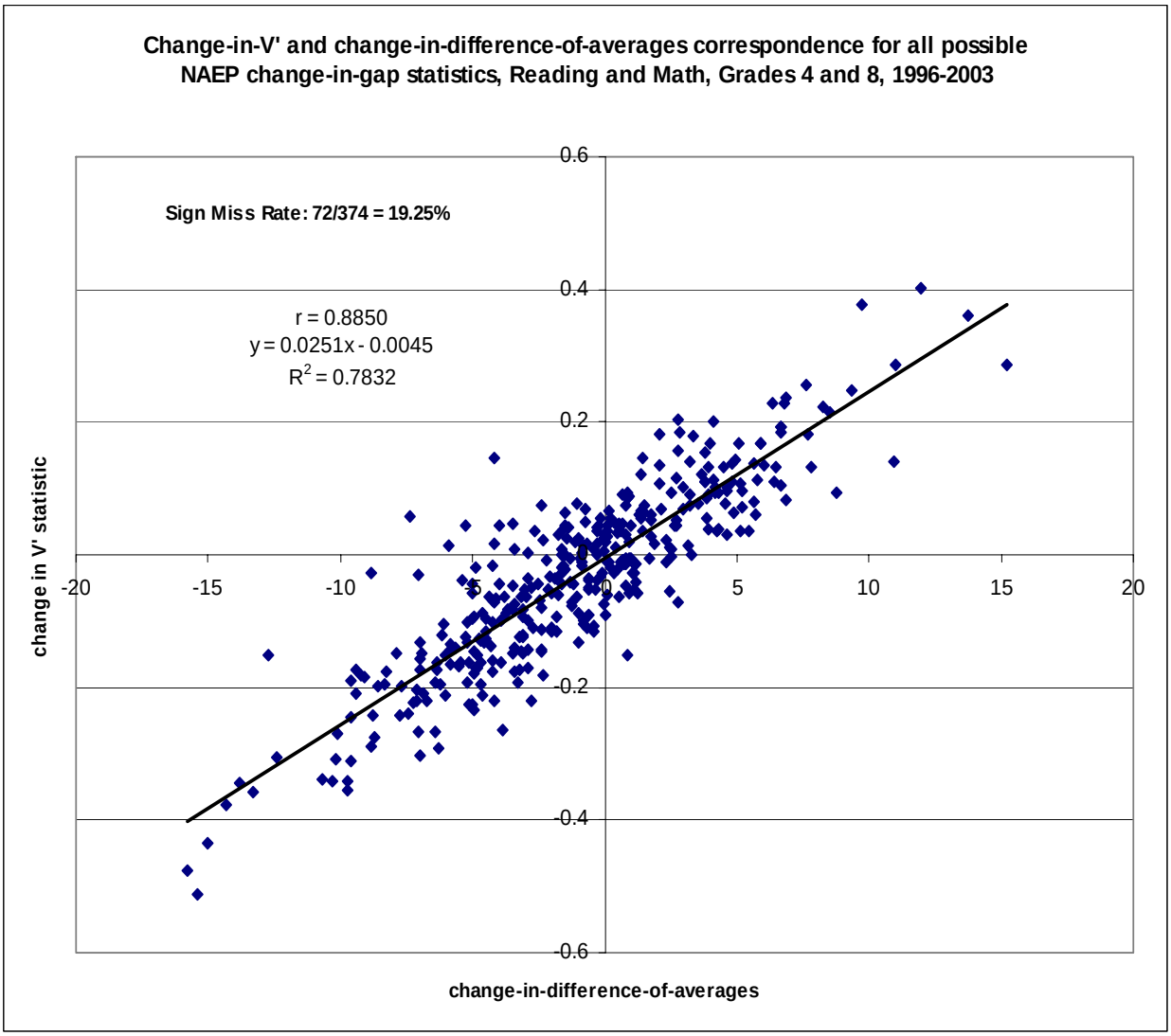


Figure 9. Change in Difference-of-Means versus Change-in-*V ′*.

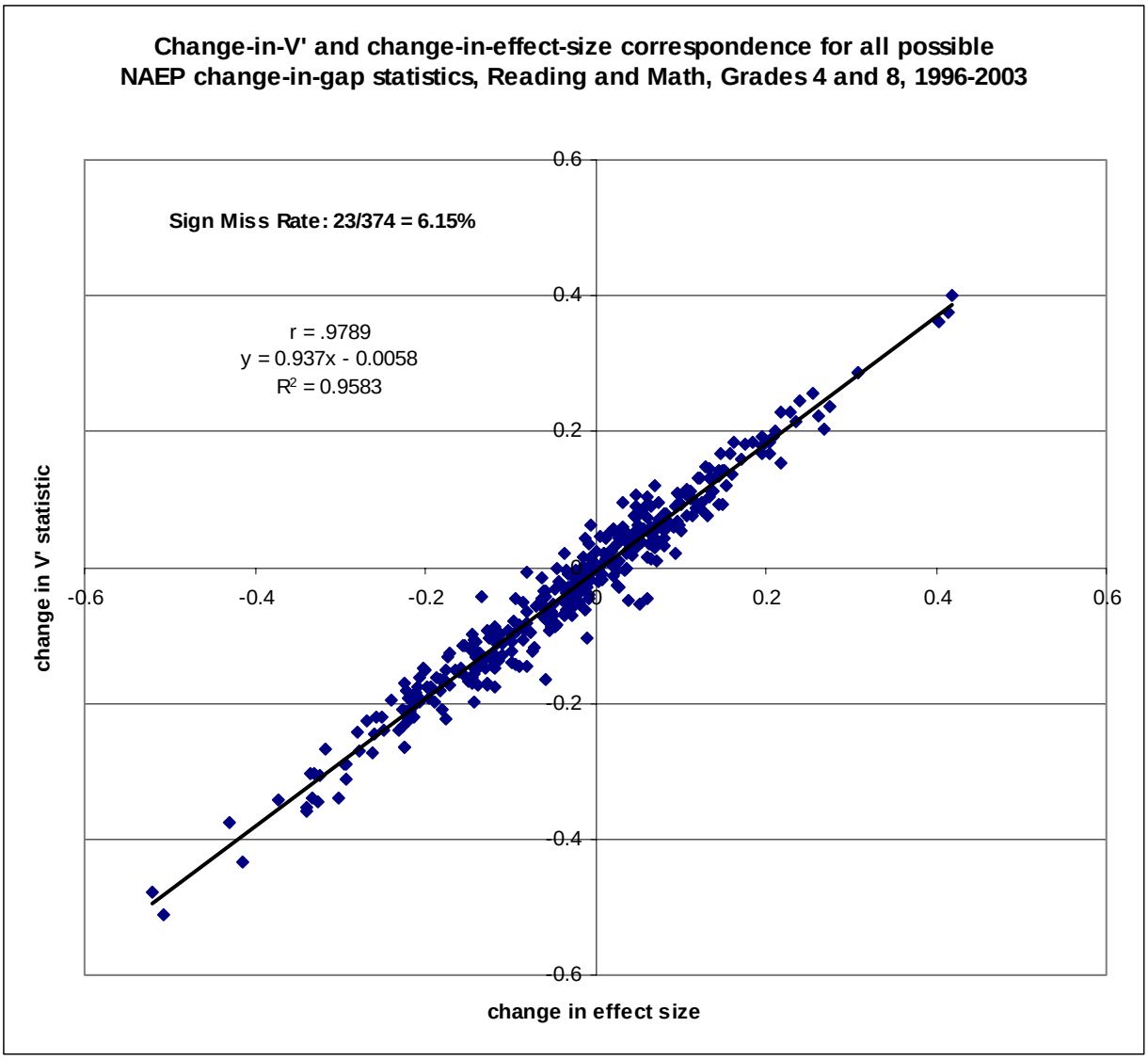


Figure 10. Change in Difference-of-Means versus Change-in-*V*.

1.5 Conclusions

This section tries to make three main points. First, average-based statistics that measure changes, gaps, and changes in gaps in test scores are made far too important by current educational policies to be dependent on the vagaries of a test score scale. Second, changes, gaps, and changes in gaps in test scores can be visualized and estimated within a metric-free framework in a robust and intuitive fashion. And third and finally, change-in-gap statistics are often susceptible to transformation-induced reversals of sign; summary statistics can mask this if graphical displays are not used as supplements in a careful analysis. While metric-

free gap statistics are useful in their own right, we believe that the most exciting applications of this metric-free framework involve the study of gap trends and, in particular, across-test comparisons of gap trends. In the growing literature comparing trends on high-stakes tests to trends on concurrent "audit" tests, these methods could be particularly useful by allowing group comparisons across the non-equivalent score scales. Results from these kinds of studies have great implications for the effects of high-stakes testing policies as they may influence educational opportunities over time. In Part II, we provide an example of this kind of analysis by using this metric-free framework to look at gap trends on four different tests over a 7-year period in California.

Part II: Hispanic-White Gap Trends on California Tests, 1998-2004

In Part I, we argued that average-based statistics, for example, the difference between the mean test score of two groups or the change in this "gap" over time, are susceptible to problems of scale. We presented a metric-free framework for evaluating gaps between test score distributions that includes both graphical and scalar-valued representations of gaps. Here, we demonstrate the use of metric-free methods in the context of representative policy questions for the state of California: How do Hispanic-White gaps on the National Assessment of Educational Progress (NAEP) for California compare to gaps on the tests administered under California's Standardized Testing and Reporting (STAR) system, and how do the gaps on these tests change over time? The data show that gaps have different magnitudes and trajectories for different testing programs. These results are consistent with the hypothesis that these tests are measuring non-identical constructs, that gains in student achievement are asymmetric with respect to these constructs, and that educational opportunities are being differentially allocated to these two groups of students.

2.1 Goodhart's Law

As Education's measurement-driven reform movement continues in its current form, the No Child Left Behind (NCLB) Act of 2001, researchers from different fields are offering cautionary anecdotes about the tendency of "accountability" policies to fail to work as intended. In particular, Goodhart's Law, a maxim posed originally in the context of monetary policy as follows, "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (Goodhart, 1975), has reached a wider audience in the Organizations and Business Management literature via generalizations like Hoskin's (1996), "every measure which becomes a target becomes a bad measure" (p. 265). As illustrations, consider what might happen if a teacher knew that her skill would be estimated solely by the grades of her students, if a surgeon knew that her merit would be estimated solely by her patients' survival rate, or if an accountant knew that her efficiency would solely be estimated by the number of tax returns she completed. In each case, the proposed measures may initially estimate true proficiencies, but as they become targets the actors will likely develop mechanisms for artificially inflating the measures that damage the measures' predictive relationships with the parameter of interest and, incidentally, that act to the detriment of the quality of services provided.

Accountability arises from trusting measures and distrusting people (Porter, 1996). NCLB places firm faith in test scores as measures of desired learning outcomes while locating responsibility for low test scores with incompetent or unmotivated educational practitioners. Goodhart's law suggests that test scores will become inflated measures, but there is little pressure to hold the measurement itself accountable. After all, accountability derives from a lack of trust, and the measure is trusted. One way to assess whether Goodhart's Law is operating in the case of NCLB policies (i.e., whether gains in test scores are being "artificially inflated," is to look at other tests that are not perceived as targets). If score trends for these "audit tests" disagree with score trends for high-stakes tests, this may suggest that Goodhart's Law is operating. In particular, we expect high-stakes test scores to rise while audit test scores show smaller gains, no gains, or a negative gain.

Of course, such a result is difficult to interpret. Under NCLB, high-stakes tests are designed to measure "academic content standards" that are set forth by each state, whereas an audit test may not have such guidelines. The evidence is just as consistent with Goodhart's Law as it is with a hypothesis that student proficiency is increasing in exactly the content areas which the state has determined are important (including those that happen not to be measured by the audit test) and not others (including other areas that may be measured by the audit test). It is also worth noting that a negative result, where trends are identical on high-stakes and audit tests, is reassuring but just as inconclusive. Strong advocates of Goodhart's Law could argue that the audit test was really just a clone of the high-stakes test, and that an ideal measure of desired learning outcomes would show discrepant trends. Discrepancies are best resolved by addressing the elements that each test intends to measure, how well each test measures these elements, and how valued these elements are by educational stakeholders.

Researchers have compared high-stakes and audit test trends previously. Koretz, McCaffrey and Hamilton (2001) introduced the terms "focal," (i.e., high-stakes, and "audit" tests in a framework for validating test score gains). A number of studies have used the National Assessment of Educational Progress (NAEP) as an audit test. As examples, Klein et al. (1998) investigated the Texas Assessment of Academic Skills (TAAS) between 1994 and 1998, Koretz and Barron (1998) investigated the Kentucky Instructional Results Information System (KIRIS), and Linn et al. (1990) reviewed a number of widely-used norm-referenced tests. None of these three studies found that gains on these tests generalized to NAEP gains. Jacob

(2002) compared Chicago public school students' gains on the Iowa Tests of Basic Skills (ITBS), a test with high stakes for students and schools, and the Illinois Goals Assessment Program (IGAP), an old high-stakes test that then had low or no stakes attached. Low achieving schools were found to have larger gains on the ITBS than on the IGAP after the stakes shifted to the ITBS.

Jacob's finding is consistent with a hypothesis set forth by McNeil (2000), who suggests that schools serving underprivileged populations will typically respond more strongly to accountability pressures. If, for underprivileged students, these large gains in high-stakes test scores do not generalize to gains on audit tests or are even associated with declines on these tests, skepticism may be cast on the framing of NCLB policies as a mechanism to improve education for the less privileged. By incorporating the principles of McNeil and Goodhart into a framework that considers score gaps between privileged and underprivileged populations, we would expect to see gaps decrease over time on high-stakes tests while decreasing less, staying the same, or increasing on audit tests.

Goodhart's Law does not necessarily make predictions about the relative magnitude of gaps on high-stakes versus audit tests; it only suggests that gap trends are likely to disagree. NCLB policies place the greatest pressures on low performing schools, where the percents of proficient students are below Annual Measurable Objectives (AMOs) set by the state. A disproportionate number of these schools have low scores on various measures of school and student socioeconomic status. If test preparation is presumed to be zero-sum, that is, if teachers dedicating time to improving performance on one test does not generalize to improving performance on another, and if test preparation was undertaken by those teachers and schools under the greatest accountability pressures, we would expect larger gaps on audit tests. On the other hand, if teaching to a high-stakes test requires specialized teaching that privileged schools can better offer whereas the audit test is more a measure of learning that goes on naturally in all classrooms, we would expect gaps on the high-stakes test to be larger. These assumptions may be compatible depending on whether a high-stakes testing program is just beginning or if it has been in effect for some time. Gaps on high-stakes tests may be larger at the beginning of a high-stakes testing program, but the relative magnitude of gaps may reverse as teachers in underprivileged schools refocus demands on high stakes tests at the expense of the untested topics within the domain.

2.2 An Overview of California Tests and Accountability Policies, 1998–2004

California passed the Public Schools Accountability Act (PSAA) in 1999. The cornerstone of the PSAA is the Academic Performance Index (API), a composite that measures the academic performance and the academic improvement of schools. The "API Base" is calculated each year for every public school; this number is used to rank schools in an absolute fashion and also with respect to schools with demographics predictive of similar achievement. The "API Growth" is calculated in the same manner as the previous year's API Base, and is used to track a school's improvement from one year to the next. For elementary and middle schools, the API also serves as an "additional indicator" for determining Adequate Yearly Progress (AYP), as required by NCLB. API Growth targets are set for each school and for each numerically significant subgroup within a school.

This section focuses only on test scores from Grades 4 and 8, as these are the two grades for which NAEP test scores have been reported for states, including California. The API has been calculated for elementary and middle schools using test scores (Grades 2–8) from three different batteries of tests: two norm-referenced test (NRT) batteries and a criterion-referenced test (CRT) battery. The two NRTs are the Stanford Achievement Test, version 9 (SAT-9) and the California Achievement Tests, 6th Edition (CAT/6), and the CRTs are the California Standards Tests (CSTs). The SAT-9 was administered in the academic year ending in 1998, but this was before the API was created. Table 3 shows how the contribution of these tests to the API changed between 1999 and 2004. Scores are reported for the SAT-9 between 1998 and 2002 (see Rogosa [2003] for related SAT-9 and API analyses), and the CAT/6 replaced the SAT-9 for 2003 and 2004. The first CST was introduced into the API in 2001 for English Language Arts (ELA). By design, it has become the most significant contributor to the API. Reading, Math, and ELA weights may not add up to 100% for any given year due to other tests like Spelling and Language. In 2004, weights for content areas varied as a function of school composition. If there are no missing data for a school, if all students have taken all tests, and if there are equal numbers of students at each grade level, then the weights are those shown in the last row of Table 3 (California Department of Education, 2005).

Table 3 also shows the years and subjects where California NAEP scores are reported. We consider NAEP as an audit test because there are no official stakes associated with NAEP test results in the state of California. In the remainder of Part II, we look at gap trends in the state of California for these four different tests, three

Table 3

An Overview of California Testing, 1998-2004

Acad. Year Ending	NRT	CRT	NAEP	API?	API NRT/CRT Weight	NRT Read.	NRT Math	CRT ELA	CRT Math
1998	SAT9		Reading	No					
1999	SAT9			Yes	100% / 0%	30%	40%		
2000	SAT9		Math	Yes	100% / 0%	30%	40%		
2001	SAT9	CST		Yes	64% / 36%	12%	40%	36%	
2002	SAT9	CST	Reading	Yes	20% / 80%	6%	8%	48%	32%
2003	CAT6	CST	Reading & Math	Yes	20% / 80%	6%	8%	48%	32%
2004	CAT6	CST		Yes	14% / 86%	4.29%	5.71%	34.29%	22.86%

high-stakes tests and one audit test. We report gaps in Grade 4-only and Grade 8-only in a cross-sectional, successive-cohort fashion. Any "improvement" that we speak of refers not to the learning of individual groups of students, but instead, for example, to the improvement of this year's fourth-grade cohort with respect to last year's fourth-grade cohort.

2.3 Methods and Data Sources

Comparing gaps and gap trends across testing programs poses the problem of comparing scores on different score scales. Test scores cannot be compared unless their scales can be transformed to a similar metric or unless the statistic to be compared is scale-independent. Scales can be "standardized" through effect size calculations (i.e., dividing the difference between two average scores by the pooled standard deviation, but these calculations are still dependent on the argument for the interval properties of the score scale). In Part I, we presented a series of gap statistics that are free of all scale information. We use these statistics here to track the difference between Hispanic and White test score distributions. The Hispanic and White classifications may be considered as rough proxies for socioeconomically disadvantaged and advantaged groups, respectively. Socioeconomic indicators like participation in free/reduced price lunch programs or Title I programs were extremely volatile in the early years of the PSAA, while racial indicators remained relatively stable. The metric-free gap statistic we choose to use is the V' statistic. The V' statistic can be calculated from the Probability-Probability (PP) Curve (Gnanadesikan, 1977). Define a cumulative distribution function (CDF) for the Hispanic and White test score distributions, $F_h(x)$ and $F_w(x)$ respectively. These CDFs return a proportion p_h or p_w representing the percentage of students from each group with a test score less than or equal to score x. The PP Plot can be defined as $p_h = F_h$

($F^{-1}_{w}(p_w)$), where the plot returns p_h as the proportion of Hispanic students at or below the p_wth percentile of the White test score distribution.

The V statistic, a statistic bounded by −1 and 1, is the area A between the PP curve and the diagonal as the proportion of the area between the diagonal and the unit square (which is 0.5 by definition). Thus, $A = \int_0^1 F_h\left(F_w^{-1}(p_w)\right) dp_w - 0.5$, and V=2A. V' is a transformation which allows the statistic to be interpreted as a kind of metric-free effect size, though the statistic is still scale-independent. The V' statistic can be interpreted as if scale information is "put back" into the metric-free statistic under the assumption that the two distributions are normally distributed with unit standard deviations. The V' statistic can thus be thought of as the gap between the two groups in standard deviation units. This transformation is achieved via equation 7.

The required NAEP data were downloaded online from the NAEP Data Tool (National Center for Educational Statistics, 2004), and all California Test Data were extracted from publicly available data sets that were either found online or obtained by contacting the California Department of Education. All publicly available California test data reported proportions above or below three cutpoints for each distribution that allowed for three PP points. By adding the theoretical points (0,0) and (1,1) and using a cubic Bezier-based interpolation function, we could approximate a continuous PP curve. The interpolation function also allowed us to numerically approximate the area under the PP curve. Simulation studies suggest that three PP points is the minimum number of points necessary for the interpolation function to obtain a reasonable approximation of the PP curve. The NAEP data offered eight points for each CDF; methods for estimating PP plots from NAEP data have been detailed in the Part I.

2.4 Results and Discussion

The results for fourth-grade Reading are shown in Figure 11. Hispanic-White test score gaps increase on NAEP between 1998 and 2003, although there is a slight drop from 1998 to 2002, followed by a sharp increase in the gap from 2002 to 2003. Gaps decrease substantially on the high-stakes NRT and CRT. The solid NRT line requires clarification, as there is a switch from the SAT-9 in 2002 to the CAT/6 in

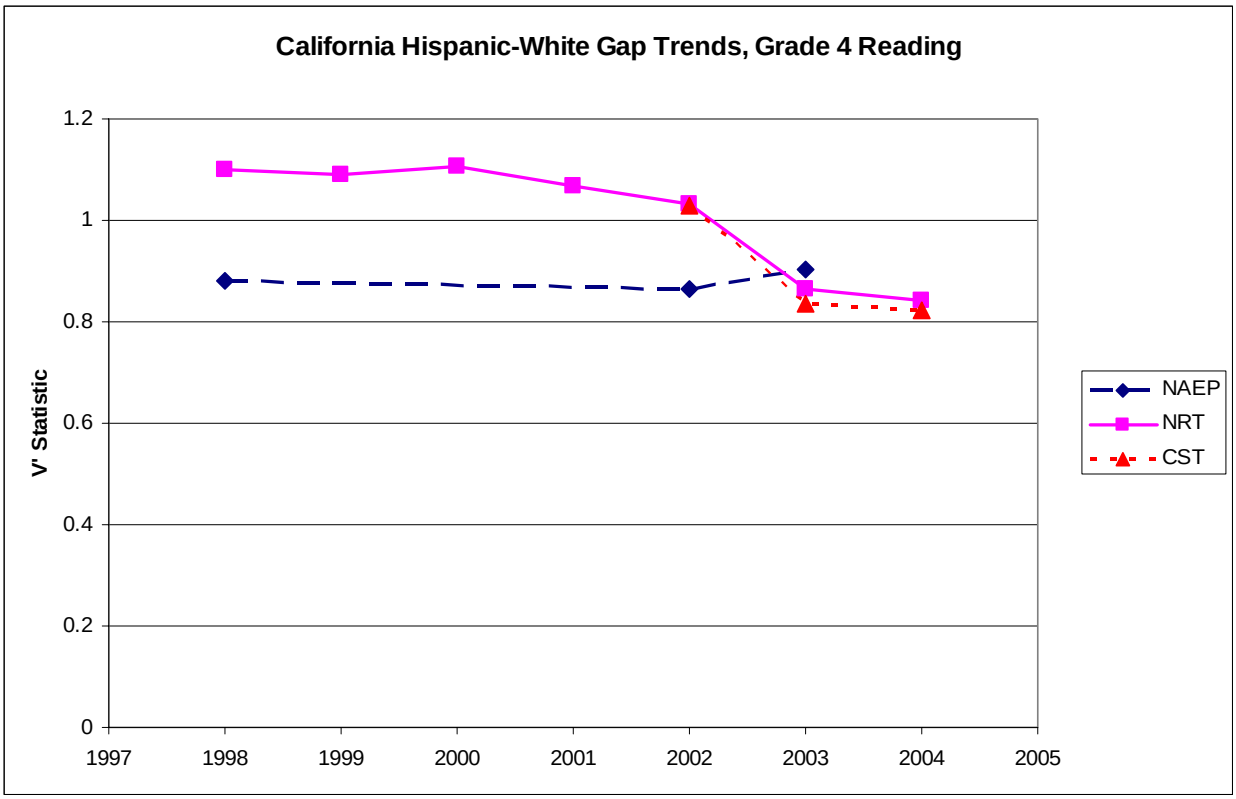


Figure 11. California Gap Trends, Grade 4 Reading.

2003. As this is a graph of score gaps, we might expect that a change in a testing program would result in an increase in the score gaps as underprivileged schools have to reallocate their limited resources to a different kind of testing practice. Instead, the decrease in the gap from 2002 to 2003 is the largest NRT gap decrease in the 7-year period. Note that changes in the norming sample from one NRT to another cannot account for any change in gaps, as the V' statistic is independent of scale.

The large decrease in the gap is consistent with the hypothesis that the 4th grade teachers and elementary schools serving predominantly Hispanic populations made a unique amount of improvement between 2002 and 2003 in teaching high-stakes proficiencies measured by both tests (relative to schools and teachers serving White students). In other words, under this hypothesis, if the SAT-9 had been administered in 2003 instead of the CAT/6, the results would have been the same. This hypothesis is also consistent with a decrease in the gap on the CST over the same time period, a decrease that is even larger than the NRT gap change. These

large decreases in gaps stand in sharp contrast to the increase in the Hispanic-White gap on NAEP between 2002 and 2003. The gap change discrepancy indicates that decreases in gaps on high-stakes tests do not tell the entire story of changing educational opportunities for disadvantaged students with respect to their more advantaged peers.

Figure 12 shows the same gap trends for fourth grade Math. NAEP data are more sparse due to unfortunate alignment of NAEP State Mathematics testing dates over this time period. The results again show gap trend discrepancies, as NAEP Hispanic-White gaps increase between 2000 and 2003 while decreasing for both high-stakes test trend lines. A notable difference between Figure 11 and Figure 12 is that audit test gaps are larger than high-stakes gaps in Figure 12 while they are smaller (until 2003) in Figure 11. This may be more attributable to the vast differences between fourth-grade Math and Reading gaps on high-stakes tests. NAEP gaps are between 0.9 and 1 for both fourth-grade Math and Reading, while high-stakes Math gaps are about 0.2 "standard deviation units" smaller than high-stakes Reading gaps. This may indicate that Hispanic students are less well

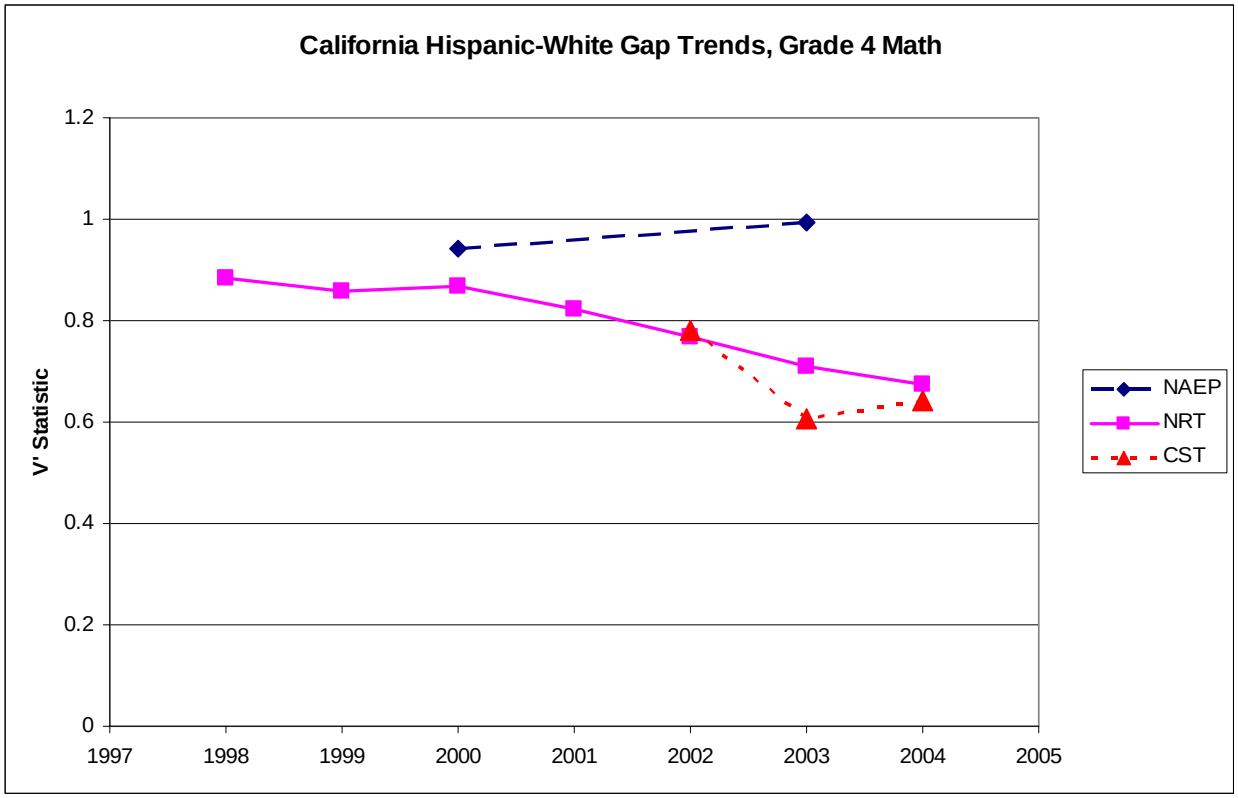


Figure 12. California Gap Trends, Grade 4 Math.

prepared for high-stakes Reading tests than they are for high-stakes Math tests. Notably, Hispanic students improve dramatically on these tasks with respect to White students, but without any similar changes in Reading proficiency as assessed

Figure 13 shows Hispanic-White gap trends for 8th Grade Reading. Gap trends for both high-stakes tests and audit tests are consistent here, and gaps on NAEP actually decrease more between 1998 and 2002 than they do on the SAT-9. Likewise, Figure 14 shows that there is gap trend consistency for Eighth Grade Math. NRT scores from 2000 to 2003 do decrease more than NAEP scores over the time period, but the discrepancy may not as politically worrisome as it would be if the sign of the trend were reversed. As was true for the Fourth Grade data, Eighth Grade gaps were larger than high-stakes gaps for NAEP Math and smaller than high-stakes gaps for NAEP Reading. This again suggests that Hispanic students are less well prepared for high-stakes Reading tests than high-stakes Math tests with respect to their White counterparts. In addition, Figure 14 shows a large discrepancy between the magnitude of gaps on the NRT and the CST, though the trends are consistent. A closer look at the actual test items of these tests may provide a defensible explanation of these discrepancies.

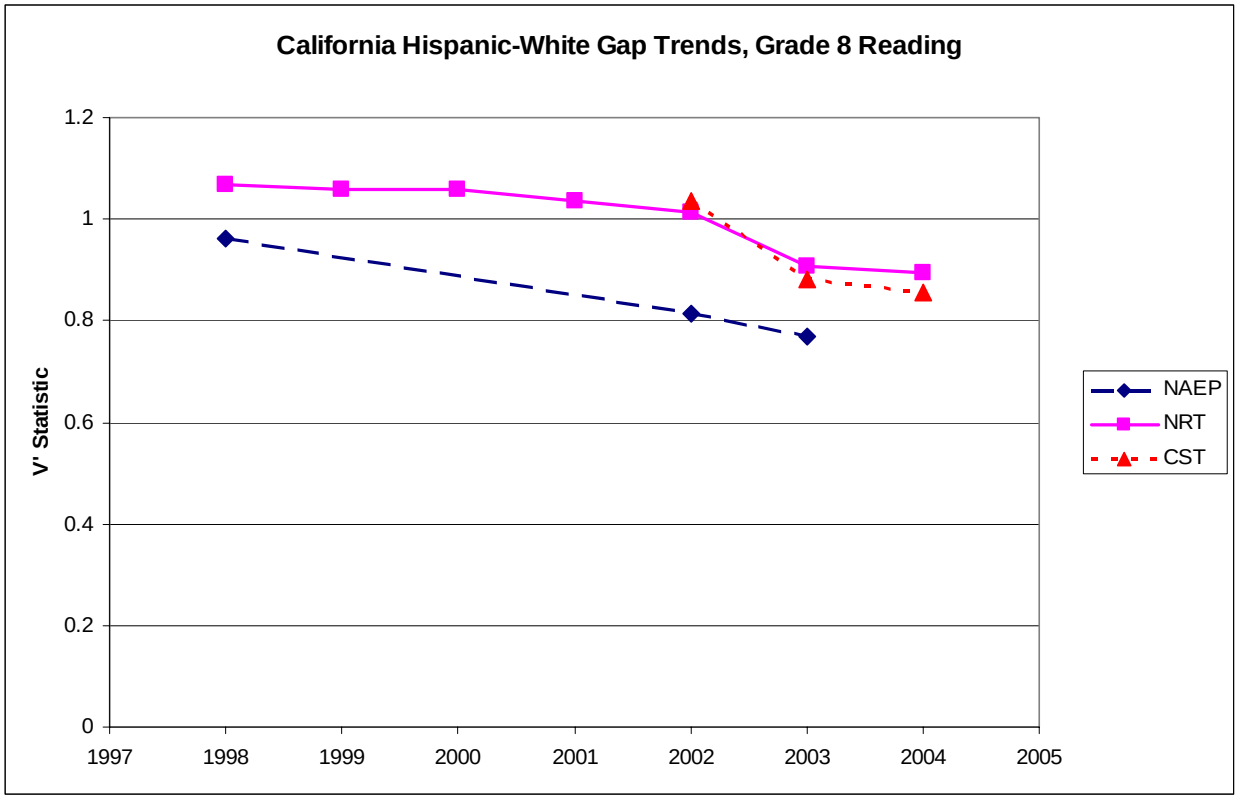


Figure 13. California Gap Trends, Grade 8 Reading.

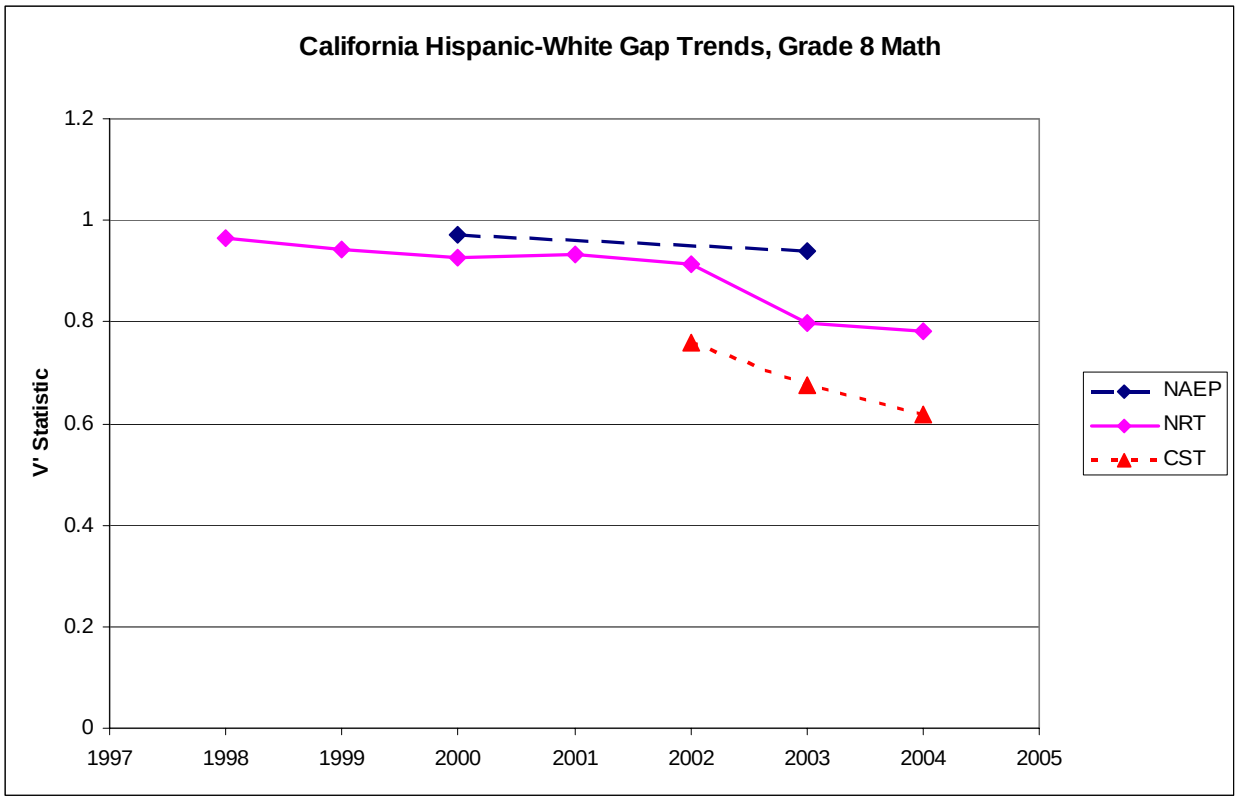


Figure 14. California Gap Trends, Grade 8 Math.

To conclude, gap trend discrepancies in line with Goodhart's Law and McNeil's accountability-response hypothesis were found in Grade 4 Math and Reading data, where Hispanic-White gaps decreased on high-stakes tests but increased on the audit test. Grade 8 Math provided marginal evidence for the hypothesis, as gaps decreased for high-stakes tests, and decreased, but to a lesser degree, on NAEP. Grade 8 Reading data did not provide any evidence for the hypothesis, as gap trends were similar between 1998 and 2003. With respect to absolute gap magnitudes, high-stakes Reading gaps were much larger than high-stakes Math gaps and NAEP gaps overall, but they showed marked improvement within the time window. These results provide evidence that gap trends for high-stakes test scores may be dangerously misleading if used as the sole indicator of the success of NCLB policies in increasing equity of outcomes. The data from Grade 4 should be of particular concern, as they show a pattern consistent with the hypothesis that high-stakes testing policies are leading to a particular kind of learning for disadvantaged students that is not generalizing to other measures of achievement.

2.5 Limitations and Next Steps

While we believe that this study raises valuable questions that policy makers should hear, it also leaves a great deal of room for improvement. To begin, the state data resolution is not entirely ideal for estimating PP plots and integrals associated with these plots. NAEP data, where eight CDF points per distribution are available, should be sufficient, but more detailed California data would be helpful in allowing a more exact estimation of V_0 statistics. Along these lines, the cubic Bezier interpolation method provides an estimate whose error has not been assessed. We are currently considering a bootstrap method of estimating the error associated with this procedure and investigating alternative methods like kernel smoothing.

School-level data would also make for a more theoretically grounded study. McNeil's hypothesis was, after all, posed at the level of individual schools' responses to accountability pressures. A hierarchical (mixed-effects) model that takes the full structure of the data into account would be more theoretically grounded and should provide better estimates of gaps and trends. This kind of analysis would be challenging to perform using NAEP as an audit test, however, because NAEP data at the school level are not publicly available. Even if nonpublic data were obtained, different samples of schools would be available for different years.

Perhaps the most logical step to take next is to open up the black boxes that these tests represent. If gaps for the Grade 8 Math CST are smaller than the gaps on the Grade 8 Math CAT/6, it makes sense to take a look at the test items from a psychological perspective to develop hypotheses about the different skills that tests require. These analyses may be well-served by multidimensional item response models and cognitively diagnostic models that take hypothesized skills required by test items into account (Ho and DiBello, 2004). Part III continues with a baseline study of trend discrepancies, where metric-free statistics are used to quantify trends instead of gaps.

Part III: Score Trend Discrepancies Between High-Stakes State Tests and the National Assessment of Educational Progress

In Parts I and II, we argued that average-based statistics, for example, a change in an average test score over time, are susceptible to problems of scale. We developed a metric-free framework for evaluating test score gaps and discussed the usefulness of this framework for comparing gaps and gap trends across tests with different score scales. Here, we show the utility of this metric-free framework as we overview test score trends in 25 states for which Reading Test data are available for both a "high-stakes" testing program and the National Assessment of Educational Progress (NAEP) in 2002 and 2003. Comparisons of metric-free measures of test score trends show that test score gains are significantly higher for State "high-stakes" Reading scores than they are for corresponding state NAEP Reading scores. These results raise serious concerns about the extent to which large-scale gains in test scores can be generalized to changes in statewide performance on broad domains such as "Reading." Disaggregation of score trends by Free and Reduced Price Lunch (FRPL) status suggests that low-income students have larger score trend discrepancies than high-income students. Metric-free statistics were calculated with limited amounts of publicly available data; more detailed studies are warranted.

3.1 An Elemental Framework

Both proponents and critics of high-stakes testing policies have a vested interest in comparing score trends on high-stakes tests to trends on similar large-scale assessments. If a high-stakes "focal" test and a low-stakes "audit" test show similar trends for similar domains (e.g., Reading), this is a reassurance that trends in the Reading performance measured by the focal test can be generalized to trends in the Reading performance measured by an audit test. Positive trends on both tests can be interpreted as convergent evidence that high-stakes testing policies are working to increase student achievement. If the trends are dissimilar (i.e., if there is a "score trend discrepancy," the messages sent to educational stakeholders become ambiguous). A particular kind of score trend discrepancy, where focal test score trends are more positive than audit test score trends, is consistent with a hypothesis that high-stakes test score trends are artificially inflated. However, this notion of "artificiality" relies on an argument for the relative validity of the audit test. To put it another way, if the focal test were a better measure of desired learning outcomes

than the audit test, attempting to "validate" focal test gains with the audit test would be misguided.

Koretz, McCaffrey and Hamilton (2001) provide a useful framework for considering how score trend discrepancies may or may not be evidence of "artificial" score inflation. They use the deliberately vague term *elements of performance* to refer to both what a test can be shown to measure and what a user may make inferences about from a test score. They define the *effective test weight* of an element as the sensitivity of a test score with respect to changes in performance on that element. However, tested elements and their weights may not always match test specifications, and elements with high effective test weights may be *intentional* or *unintentional*. Users have *targets of inference* that are composed of elements of performance, and they also have a *model of gains* that describes how a change in test scores should be consistent with changes in performance on these elements. Koretz et al. acknowledge that these targets and models are often tacit and poorly formed. Just as tested elements have test weights, elements that are part of a user's target of inference have *inference weights*, where a large inference weight for a user on a particular element refers to that user ascribing a high value to that particular element within the domain.

It is useful to consider the oft-cited "teaching to the test" hypothesis in this framework. This hypothesis holds that when high stakes are attached to a test, actions taken to improve scores on that test render it less accurate as an indicator of underlying achievement. Some such actions include a) an increased focus on teaching those content elements thought likely to be tested, at the expense of other elements in the domain the test is intended to represent, b) a focus on demonstrating achievement in just the manner called for on the test, and c) teaching of ancillary "test wiseness" skills, like time management or intelligent guessing, aimed at improving test scores without affecting underlying knowledge. These actions correspond loosely to the terms reallocation, alignment, and coaching from Koretz et al. Though "teaching to the test" almost always has a negative connotation, differing targets of inference may allow users to view stakes-inspired test preparation positively or negatively. If the audit test samples a different set of content elements from those on the high-stakes test, or if it frames questions in different ways or uses different item formats (i.e., differing *non-substantive elements*), then high-stakes score gains attributable to "teaching to the test" will not be fully reflected in audit test gains.

We can call this explanation of score trend discrepancies an "elemental discrepancy" hypothesis. That is, the focal and audit tests have different tested elements. In addition, it is possible in some cases that some students thought likely to earn low scores on a focal test may be excluded in one way or another from the group tested. If the audit test samples these students, or especially if sampling patterns are inconsistent across tests and from one year to the next, score trend discrepancies may result. We can call this explanation a "sampling discrepancy" hypothesis. Under this hypothesis, even if the two tests are designed to the same specifications, score trend discrepancies may result due to differing sampling schemes. Both sampling discrepancy and elemental discrepancy hypotheses may explain score trend discrepancies between the focal and audit tests overviewed in Part III[3].

3.2 Plausible and Implausible Elemental and Sampling Discrepancy Hypotheses

This section's focal tests are large-scale state Reading tests that are part of state school accountability policies. They are all high-stakes because of the sanctions for schools and districts that fail to meet requirements for Adequate Yearly Progress (AYP) mandated under the federal No Child Left Behind (NCLB) Act of 2001. The audit test is the National Assessment of Educational Progress (NAEP) in Reading. Due to the varying names of state testing programs and Reading tests within these programs, we will call all focal tests "State" with the capital; these are contrasted with NAEP. NAEP Reading results are available for most states in 2002 and 2003; this is the time period we will investigate.

NCLB policies discourage two practices that may have led to "sampling discrepancies" in the past. First, in order to receive Title I funding, districts must agree to participate in NAEP if they are drawn as part of a state sample. Thus, bias due to nonresponse at the school level is greatly reduced. Previously, in spite of the use of replacement schools, weighting adjustments, and minimum required participation levels for reportable results, a score trend discrepancy may have arisen from a number of schools opting out of NAEP at one time point or another. NAEP has a school substitution procedure in place to minimize nonresponse bias, and school participation should be 100% from 2003 on.

[3] A third discrepancy hypothesis involves differing changes in student motivation over time. This list of discrepancy hypotheses is not intended to be exhaustive.

A second practice discouraged by NCLB policies is the exclusion of low-performing students from score summaries. An increase in the exclusion rates for low-performing students would lead to an artificial inflation in test scores, and differing trends in exclusion rates between NAEP and State tests may account for score trend discrepancies. NCLB mandates a 95% participation rate for students in all schools and for each numerically significant subgroup within a school. Depending on the length of time it took for NCLB policies to phase in after being signed into law in early 2002, exclusion rates for low-performing students may have been high in 2002 and then decreased under NCLB policies in 2003. If this were the case, we would expect State test scores to show deflated trends.

Finally, it is important to remember that the trends we are reporting are cross-sectional. The fourth graders whose performance we are measuring in 2002 are not the same fourth graders whose performance we are measuring in 2003. If there are demographic shifts in the population or changes in out-of-school factors, test scores may be influenced as a result. Most of these changes are unlikely to explain score trend discrepancies, but there are certainly plausible scenarios. A large change in the proportion of English Learners may differentially impact NAEP and State Reading test scores in accordance with the different Reading skills each test measures. Alternatively, an out-of-school program may be initiated which privileges the learning of skills disproportionately measured by one test. These changes would only lead to a score trend *discrepancy* if the tested elements were nonidentical. Thus, these particular "elemental discrepancy" hypotheses are potential alternatives to the "teaching to the test" elemental discrepancy hypothesis.

3.3 Socioeconomic Resources and Score Trend Discrepancies

The "teaching to the test" hypothesis assumes that there are significant incentives to teach a narrowly defined curriculum. These incentives are clearly not equal across all schools and classrooms. A school that has a high percentage of students scoring at "proficient" or above may have little reason to change current practices. Schools that have limited resources may have to devote all available resources to meeting the benchmarks set by NCLB policies. We might expect students in these schools and classrooms to show the most evidence of "learning to the test," and thus show the greatest score trend discrepancies between that test and an audit test.

In upcoming sections, we use participation in the National School Lunch Program as a rough proxy for student socioeconomic status. We disaggregate score trend discrepancies for students who are eligible and ineligible for these programs to investigate the hypothesis that accountability policies may have a disproportionate impact on students and teachers in low-resource schools.

3.4 Metric-Free Methods

Score trends are usually expressed in one of two ways. The first is to take the difference between the mean scaled score (MSS) at time 2 and the MSS at time 1. The second is to take the percent of examinees above some cut point (percent above cut, or PAC) on a score scale and take the difference between the PAC at time 2 and the PAC at time 1. The second expression has come into widespread use due to NCLB, because Annual Measurable Objectives (AMOs) defining Adequate Yearly Progress (AYP) are stated in terms of a required percent "proficient" or above. Positive values for either of these expressions are interpreted as showing increases in student achievement.

Both of these expressions have technical drawbacks. Part I reviews studies and shows examples demonstrating that the rank order of mean scores may not be meaningful if the interval properties of the score scale are suspect. In technical terms, we can only say that one mean score will be greater than another under any positive monotone transformation of the score scale if the two distributions are "stochastically ordered." If this condition does not hold, an increase in average student achievement may become a decrease in average student achievement under a plausible transformation of the test score scale. Holland (2002) shows us that interpretations of changes in PAC values (and a fortiori changes in gaps measured by PAC values) may likewise be dependent on an arbitrary consideration, in this case the choice of the "proficiency" cut score.

These technical issues are compounded by a third issue when we become interested in comparing focal and audit test scores. MSS difference expressions become impossible to compare because focal and audit score scales are different. A common solution is to divide the MSS difference expressions by the pooled standard deviations at the two time points. This is a solution that does not address the pliability of either of the score scales in question. PAC values are even more difficult to compare because standard setting procedures may be inconsistent across tests. If PAC difference values are already dependent on the location of the cut score,

comparing difference values across tests with differing cut scores, possibly established by different methods, may be misleading.

To overcome the problems associated with noncommensurable test score scales and arbitrarily established cut scores, Part I argues for a metric-free, graphical representation of test score trends. When a scalar value is required, the V' statistic is recommended, which can be interpreted loosely as a "metric-free effect size" statistic. This statistic does not show whether distributions are stochastically ordered at a glance, but, in a sense, whether a mean value takes on a positive or negative value under arbitrary transformations becomes an auxiliary concern given that the V' statistic can be shown to be positive or negative under all transformations. The remainder of this section compares test score trends on NAEP and State Reading tests from the time period 2002 to 2003. These trends will be expressed as V' statistics.

3.5 Data Availability

Under NCLB policies, starting in 2003, all states must participate in biannual NAEP Reading and Mathematics assessments at the fourth- and eighth-grade levels as a condition for receipt of Title I funding. In addition, school districts that receive Title I funding must agree to participate in NAEP if asked to do so. As a result, NAEP Reading data are available for all states in 2003. In 2002, however, seven states did not participate in NAEP, and two more states had data available for fourth but not eighth graders. State test data were gathered by visiting state websites and downloading applicable data. State data may not have been available or useful for one or more of the following reasons.

- The state does not publish appropriate data online from 2002, 2003, or both years.
- The state does not test Reading in Grade 4 and/or Grade 8.
- The state is using a different testing program in 2003 than it did in 2002.
- The state does not test in the spring, thus its testing is not aligned with NAEP testing.

States for which spring 2002 and 2003 data are available for both State and NAEP Reading tests are shown in Table 4. Data availability for students who participate in the National Free and Reduced Price Lunch (FRPL) program is also

State	Grade 4	Grade 8	FRPL 4	FRPL 8	NFRPL 4	NFRPL 8
Arizona		3				
California	1	1	1	1		
Connecticut	4	4	4	4	4	4
Delaware		4		4		4
Georgia	2	2				
Hawaii		3		3		
Kansas		1		1		1
Kentucky	7		3		3	
Louisiana	4	4				
Maine	3	3				
Massachusetts	3					
Mississippi	2	2	2	2	2	2
Montana	3	3	3	3	3	3
North Carolina	3	3	1	1	1	1
Ohio	3					
Oklahoma		3		1		
Oregon		1				
Pennsylvania		3		3		
Rhode Island	4	4				
South Carolina	3	3	3	3	3	3
Utah	3	3				
Vermont	3	3	3	3	3	3
Virginia		2		2		
Washington	3					
Wyoming	3	3	2	2	1	1
COUNT 25	17	21	9	14	8	9

Table 4. State PP Cutpoint Availability, 2002-2003.

indicated, along with data availability for students who are not in the program (NFRPL). The entries in the table show the number of data points available to estimate the metric-free trend statistic; this will be elaborated upon in the next section. If there is no entry, no data were available for one or more of the reasons listed above.

The number of states that had appropriate data available for each category for both the state and NAEP tests is listed in the last row. Twenty-five states had at least one estimable metric-free trend comparison. In general, Grade 8 data were more readily available than Grade 4 data. For various reasons, possibly related to the history of grade-subject combinations in State NAEP, many states tested Reading but not Mathematics in Grade 8 and Mathematics but not Reading in Grade 4.

3.6 Results

Statewide score trend discrepancies are shown in Figure 15 for Grades 4 and 8. State test trends are on the x-axis; NAEP trends are on the y-axis; and positive trends indicate an increase in student achievement. All trends shown are V' statistics and can be interpreted as metric-free effect sizes. Identical trends should line up on the diagonal. Any point below the diagonal is an example of a state-grade combination where State test trends are more positive than NAEP trends. The 38 data points represent 25 states for which State and NAEP trends could be estimated. Thirteen of the states had estimable trends for both Grades 4 and 8, and 12 states had an estimable trend for only one of Grades 4 and 8.

Twenty-five of the 38 data points, or approximately 2 out of 3 state-grade combinations, have State test trends that are more positive than NAEP test trends. A matched sample t-test shows that State tests trends are, on average, significantly more positive than their NAEP counterparts ($t = 3.719$, $df = 37$, $p < .001$). These averages are shown as an open circle in the fourth quadrant, representing an average State gain of .034 and an average NAEP decline of −.028. The four data points in quadrant 2 and the 15 data points in quadrant 4 have notable political significance because they lead to dramatically ambiguous policy conclusions, where it appears that Reading achievement is both improving and declining. The relatively large number of data points below the diagonal, and especially the number of data points in quadrant 4, are consistent with a "teaching to the test" hypothesis.

Figure 16 shows statewide score trends disaggregated by eligibility for the National School Lunch Program. Fifteen states had at least one estimable disaggregated trend for both State and NAEP tests. Four trend discrepancies are possible for each state, one for each grade, and one each for eligible (FRPL) and ineligible (NFRPL) students. Forty trend discrepancies are plotted in Figure 16; 23 for FRPL students and 17 for NFRPL students. An overall matched sample t-test shows that the average State trend is significantly higher than the average NAEP trend ($t = 3.633$, $df = 39$, $p < .001$).

Twenty out of 23 (87%) FRPL state-grade combinations are below the diagonal, whereas only about half (9 of 17, 53%) of NFRPL state-grade combinations are below the diagonal. For NFRPL students, average NAEP trends are not as positive as State trends, but the difference is not significant ($t = 1.325$, $df = 16$, $p \approx .20$). For FRPL students, average State trends are significantly more positive than NAEP trends

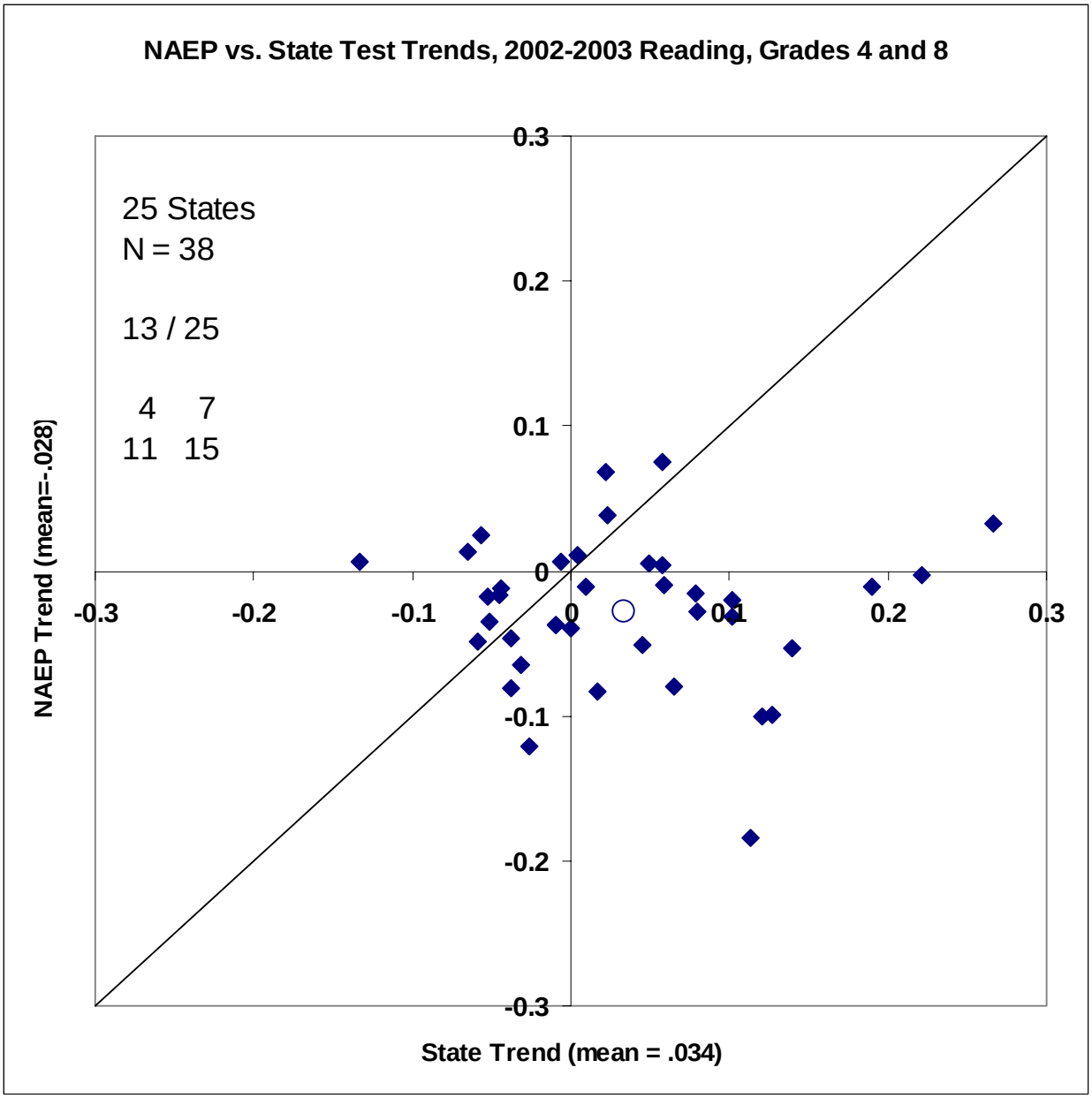


Figure 15. Score Trend Discrepancies, 2002-2003 Reading.

(t = 3.935, df = 22, p < .001). As shown by the FRPL and NFRPL mean trend discrepancies on the graph, average FRPL trend discrepancies are greater than NFRPL trend discrepancies. However, using only the 17 state-grade combinations that have both FRPL and NFRPL estimates of trend discrepancy, we find that FRPL trend discrepancies are not significantly greater than NFRPL trend discrepancies (t = 1.275, df = 16, p≈ .22).

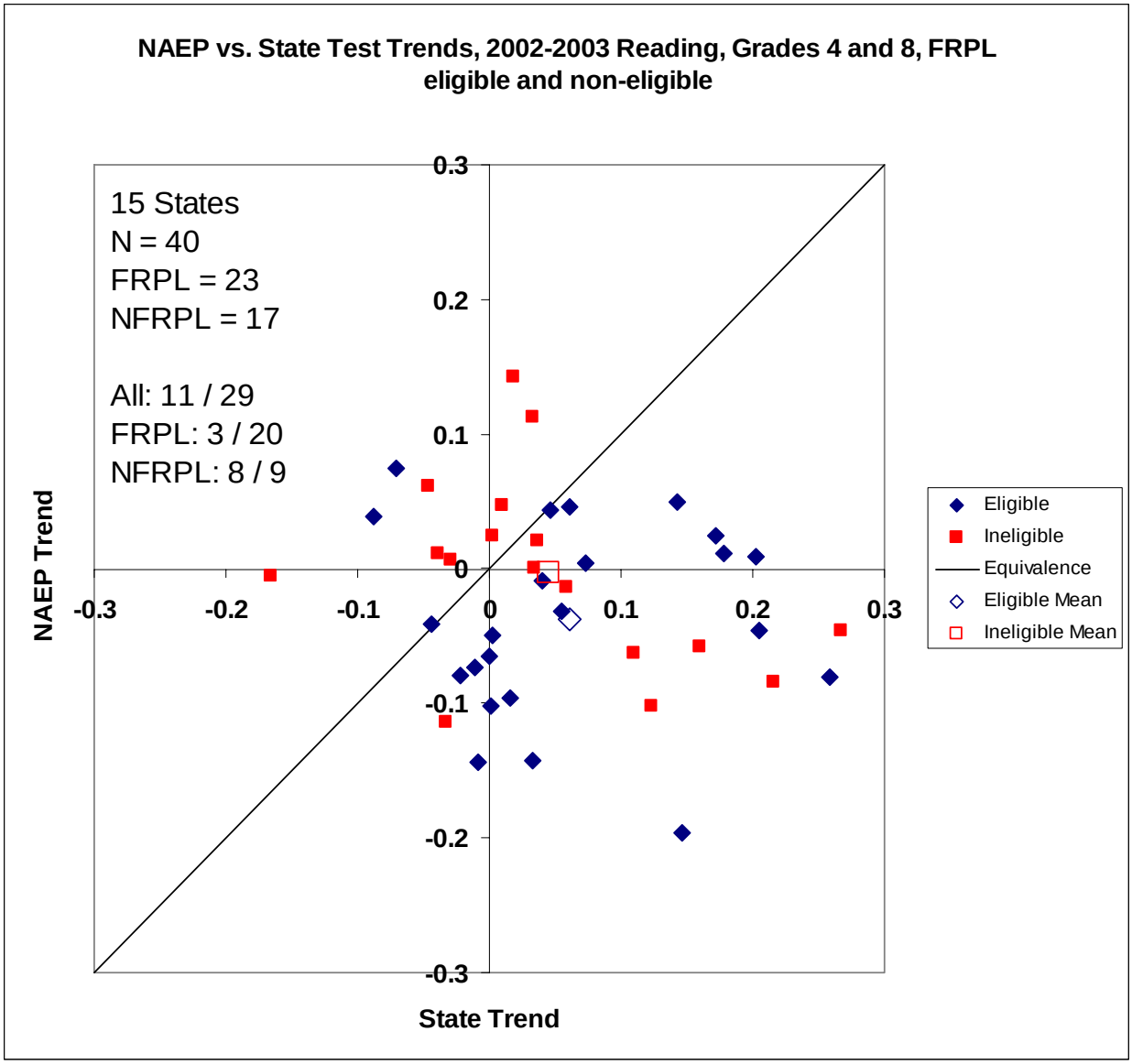


Figure 16. Score Trend Discrepancies by FRPL Eligibility.

3.7 Discussion

Focusing on statistical significance by averaging over states and grades may understate the importance of each of the data points in Figures 15 and 16. Any data point off the diagonal is, in itself, a noteworthy score trend discrepancy with three caveats. The first caveat is that V' statistics may be masking a lack of stochastic ordering, and that PP plots that represent the difference between the two distributions may in fact be crisscrossing the diagonal. If there is no stochastic ordering, then, for example, the number of students at or above Proficient may

increase while the number of students at or above Advanced may decrease. This is equivalent to stating that transformations exist that can flip the ranking of the means at time 1 and time 2. In these cases, it becomes more difficult to say precisely what we mean when we make the claim that trends in student performance are either positive or negative, and it would become more difficult to compare these ambiguous trends across tests.

The second caveat is that NAEP estimates of percentiles and PAC measures have notable standard errors that lead to sampling variability in the NAEP V' estimates. Finally, V' statistics are estimated from limited amounts of publicly available data. A PP plot that represents the gap between two distributions may be estimated from as few as 2 Proportion-Proportion pairs, where each pair represents the proportions at or below a given cut point for both distributions. The number of PP pairs for each V' estimate are the entries in Table 15. In cases where only one PP pair is available, an effect size statistic is estimated assuming that the two distributions are normal with unit variance. The accuracy of these estimation methods is still an area of ongoing research, but preliminary studies suggest that, if anything, trends in conditions of limited data are underestimated; thus, trend discrepancies with full data availability will in most cases be even greater.

It should be of significant concern to policy makers and educational stakeholders that average State test gains are considerably greater than average NAEP gains. This concern should be magnified by the large numbers of state-grade combinations that show not only trend discrepancies but trend sign discrepancies. If these discrepancies can be attributed to systematic differences between the populations framed for State versus NAEP testing, then they should be carefully investigated. If gains or declines on test scores on either NAEP or State tests can merely be attributed to sampling fluctuations, then widespread interpretations of these trends as improvement or decline in educational achievement must be approached with greater caution. If these discrepancies can be attributed content discrepancies, the cognitive domain of these tests needs to be mapped, and trends on overlapping and nonoverlapping content areas need to be estimated. If overlapping content areas can show the same trends and nonoverlapping content areas can account for score trend discrepancies, a more coherent picture of changes in educational achievement and school success may emerge.

References

California Department of Education. (2005). 2004 Academic Performance Index Base Report: Information guide.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: John Wiley & Sons.

Goodhart, C. (1975). Monetary relationships: A view from Threadneedle Street. In *Papers in Monetary Economics*, volume 1. Reserve Bank of Australia.

Haertel, E., Thrash, W., & Wiley, D. (1978). Metric-free distributional comparisons. Chicago: ML-Group for Policy Studies in Education.

Ho, A., & DiBello, L. (2004). Policy applications of the fusion model for skills diagnosis. Paper presented at the 2004 Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, *27*, 1, 3–17.

Hoskin, K. (1996). The 'awful idea of accountability': Inscribing people into the measurement of objects. In Munro, R. & Mouritsen, J., editors, *Accountability: Power, Ethos and the Technologies of Managing*, pages 265–282. London: International Thomson Business Press.

Jacob, B. (2002). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. Working Paper #8968, National Bureau of Economic Research (NBER), Cambridge, MA.

Kentucky Department of Education (2004). 2004 CATS interpretive guide: Detailed information on using your score reports. Version 2.01.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (1998). What do test scores in Texas tell us? Santa Monica, CA: RAND.

Koretz, D., & Barron, S. (1998). The validity of gains on the Kentucky Instructional Results Information System (KIRIS). Report, Santa Monica, CA: RAND.

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). Toward a framework for validating gains under high-stakes conditions. CSE Tech. Rep. No. 551, Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Lehmann, E. (1955). Ordered families of distributions. *Annals of Mathematical Statistics*, *26,* 399–419.

Linn, R., Graue, M., & Sanders, N. (1990). Comparing state and district results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice*, *9*(3), 5–14.

Lord, F. (1975). The "ability" scale in item characteristic curve theory. *Psychometrika*, *20,* 299–326.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

McNeil, L. (2000). *Contradictions of school reform: The educational costs of standardized testing*. New York: Routledge.

National Center for Educational Statistics. (2004). NAEP Data Tool. Data Retrieved between November and December, 2004.

Porter, T. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Public Law #107-110 (2002). The No Child Left Behind Act of 2001. 115 Stat. 1425.

Rogosa, D. (2003). Four-peat: Data analysis results from uncharacteristic continuity in California student testing programs. Stanford, CA: California Department of Education.

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

Tukey, J. (1977). *Exploratory Data Analysis*. Cambridge, MA: Addison-Wesley.

Yen, W. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23,* 299–326.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Measurement*, *20,* 299–326.