James P. Scanlan
Attorney at Law
1529 Wisconsin Avenue, NW
Washington, D.C. 20007
(202) 338-9224
jps@jpscanlan.com

April 30, 2014

David Britt
Chair of the Board of Directors
Kati Haycock
President and Member of the Board of Directors
The Education Trust
1250 H Street, NW
Washington, DC 20005

Re: Measurement Issues Pertaining to Education Trust Research

Dear Chair Britt and President Haycock:

On occasion I write to institutions whose activities involve the interpretation of data on demographic differences in the law or the social and medical sciences alerting them to ways in which their activities are undermined by the failure to recognize patterns by which standard measures of differences between favorable or adverse outcome rates of advantaged and disadvantaged groups tend to be systematically affected by the overall prevalence of an outcome. Other recipients of letters involving the statistical issues discussed in this letter include Robert Wood Johnson Foundation [1](Apr. 8, 2009), National Quality Forum (Oct. 22, 2009), Institute of Medicine (June 1, 2010), The Commonwealth Fund (June 1, 2010), United States Department of Education (Apr. 18, 2012), United States Department of Justice (Apr. 23, 2012), Federal Reserve Board (March 4, 2013), Harvard University (Oct. 9, 2012), Harvard Medical School and Massachusetts General Hospital (Oct. 26, 2012), Senate Committee on Health, Education, Labor and Pensions (Apr. 1, 2013), Mailman School of Public Health of Columbia University (May 24, 2013), and the Investigations and Oversight Subcommittee of House Finance Committee (Dec. 4, 2013).

This letter is most directly prompted by the Education Trust's April 2, 2014 issuance of a report titled "Falling Out of the Lead: Following High Achievers Through High School and Beyond." The report was the second of a series examining demographic differences in reaching certain levels of academic success with a focus on higher-achieving students, the first of which was a May 14, 2013 report titled "Breaking the Glass Ceiling of Achievement for Low-Income

[1] To facilitate consideration of the issues raised in letters such as this I make available electronic copies of the letters on the Institutional Correspondence subpage of the Measuring Health Disparities page of jpscanlan.com. Underlinings in this letter reflect links to the underlined material in such a copy of the letter. If the letter is corrected after it is first posted on the website, such fact will be noted on the final page.

Students and Students of Color." Both are now treated on web pages of jpscanlan.com, which I discuss below, after first explaining the pertinent statistical issues.

For reasons related to the shapes of the normal risk distributions, all standard measures of differences between outcome rates of advantaged and disadvantaged groups tend to be systematically affected by the overall prevalence of an outcome. Most notably, the rarer an outcome, the greater tends to be the relative difference in experiencing it and the smaller tends to be the relative difference in avoiding it. Thus, for example, lowering a test cutoff (or generally improving test performance), while tending to reduce relative differences in pass rates, will tend to increase relative difference in failure rates. Numerous discussions of this pattern and implications of the failure to understand it in particular contexts may be found on the pages of jpscanlan.com devoted to measurement issues (as well as in the letters referenced above).

Three recent, relatively succinct explanations of these patterns, using test score data for demonstrative purposes, may be found in "Things government doesn't know about racial disparities," *The Hill* (Jan. 28, 2014), "The Paradox of Lowering Standards," *Baltimore Sun* (Aug. 5, 2013), and "Misunderstanding of Statistics Leads to Misguided Law Enforcement Policies, " *Amstat News* (Dec. 2012). Those three articles also explain that, contrary to view of the United States Departments of Education and Justice, generally reducing public school suspension and expulsion rates, while tending to reduce relative racial/ethnic differences in rates of avoiding those outcomes, will tend to increase relative racial/ethnic differences in suspension and expulsion rates.[2] More elaborate treatments of the patterns whereby the two relative differences tend to change in opposite directions as the prevalence of an outcome changes, also using test score data for demonstrative purposes and including many graphical and tabular illustrations, may be found in my November 2013 Federal Committee on Statistical Methodology 2013 Research Conference paper titled "Measuring Health and Healthcare Disparities" (FCSM Paper), my September 2013 University of Kansas School of Law Faculty Workshop paper titled "The Mismeasure of Discrimination," and my October 2012 Applied Statistics Workshop at Harvard's Institute for Quantitative Social Science titled "The Mismeasure of Group Differences in the Law and the Social and Medical Sciences."

Table 1 illustrates the patterns described in the three articles discussed above whereby lowering a test cutoff tends to reduce relative differences in pass rates while increasing relative differences in failure rates. The table is based on a situation where two groups have normal test score distributions with means that differ by half a standard deviation (and where the standard deviations of the distributions are equal). The table presents the pass and fail rates for the advantaged group (AG) and the disadvantaged group (DG) at two different cutoff points. It also shows the ratio of DG's failure rate to AG's failure rate and the ratio of AG's pass rate to DG's

[2] More than two dozen other articles explaining these patterns in various contexts may be found on the Bibliography subpage of the Scanlan's Rule page of jpscanlan.com. The more extensive of these include "Race and Mortality Revisited," *Society* (July/Aug. 2014) (forthcoming), "Can We Actually Measure Health Disparities?" *Chance* (Spring 2006), "Race and Mortality, *Society* (Jan/Feb 2000), "Divining Difference," *Chance* (Fall 1994), "The Perils of Provocative Statistics," *Public Interest* (Winter 1991).

pass rates at the two cutoffs. The table thus shows how lowering the cutoff (or improving test performance such as to enable everyone scoring between the two points now to pass the test) – and thereby making test failure less common and test passage more common – tends to increase the relative difference in failure rates while reducing the relative difference in pass rates. [3] The final column of the table also presents the absolute (percentage point) difference between the pass (or failure) rates at each cutoff, but I will briefly defer discussion of the absolute difference.

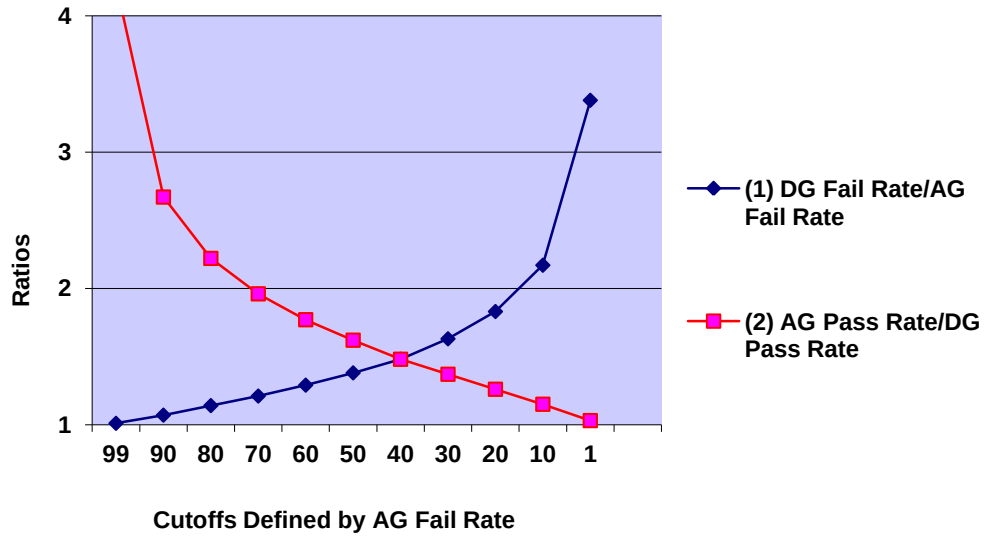
Table 1. Pass and Fail rates of Advantaged Group (AG) and Disadvantaged Group (DG) at Two Cutoffs, with Ratios of (1) DG Fail Rate to AG Fail Rate and (2) AG Pass Rate to DG Pass Rate at Various Cutoff Points Defined by AG Fail Rate, and Absolute Difference between Rates

Cutoff	AG Pass	DG Pass	AG Fail	DG Fail	DG/AG Fail Ratio	AG/DG Pass Ratio	Absolute Difference
High	80%	63%	20%	37%	1.85	1.27	0.17
Low	95%	87%	5%	13%	2.60	1.09	0.08

Figure 1 illustrates the same pattern across the entire range of test scores. The numbers at the bottom of the figure are the failure rates of AG, which are used as benchmarks for the overall prevalence of test failure. The line with the diamond marker (red in the electronic version of this letter) tracks the ratio of DG's failure rate to AG's failure rate, and the line with the square marker (blue in the electronic version) tracks the ratio of AG's pass rate to DG's pass rate, at each benchmark. From left to right, the lines illustrate the effects on the two ratios of serially lowering the test cutoff from a point where almost everyone fails to a point where almost everyone passes, in each instance enabling all persons with scores above each new cutoff now to pass the test. The figure thus illustrates the common pattern whereby, as the prevalence of an outcome changes, relative differences in experiencing it and relative differences in avoiding it tend to change in opposite directions.

[3] The ratio is commonly termed the "rate ratio, "risk ratio," or "relative risk" (RR). The relative difference between rates is RR minus 1 where RR is greater than 1 (in which case the larger the RR the larger the relative difference) and 1 minus RR where RR is less than 1 (in which case the smaller the RR the larger the relative difference). In recent years I have generally used the larger figure as the numerator of the RR for both favorable and adverse outcomes. Thus, as to both outcomes, the larger the RR the larger the relative difference. Whether one uses the larger or smaller figure as the numerator in RR can affect the size of a relative difference. For example, in a case where rates are 30 percent and 40 percent, the former could be deemed 25 percent less than the latter or the latter could be deemed 33 percent greater than the former. But choice of numerator is irrelevant to issues about the comparative sizes of relative differences addressed here. Determinations as to which is the larger relative difference reflected by two pairs of rates of experiencing an outcome will always hold regardless of which figure is used as the numerator of the ratio.

Figure 1. Ratios of (1) DG Fail Rate to AG Fail Rate and (2) AG Pass Rate to DG Pass Rate at Various Cutoff Points Defined by AG Fail Rate

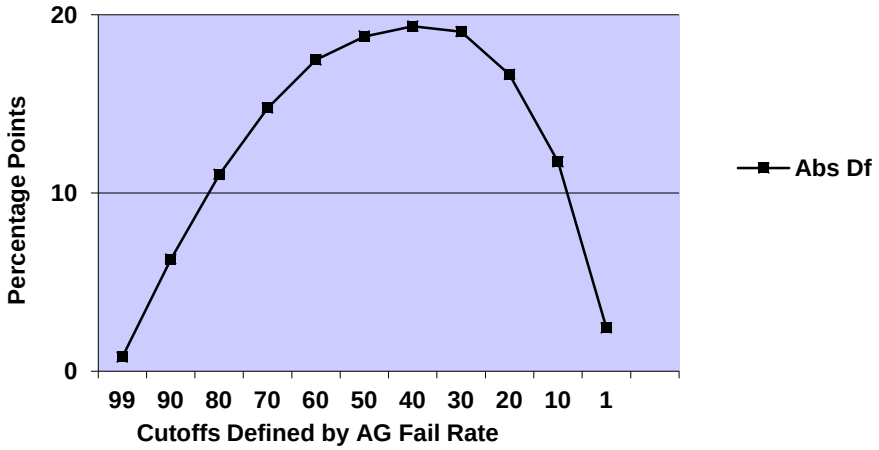


The absolute difference between rates – 17 percentage points with the higher cutoff and 8 percentage points with the lower cutoff in Table 1 – is unaffected by whether one examines the favorable outcome or the corresponding adverse outcome. But for a measure to effectively quantify the forces causing the outcome rates of two groups to differ (or, put another way, the difference in the circumstances of two groups reflected by their differing outcome rates), the measure must remain constant when there occurs an overall change in the prevalence of an outcome akin to that effected by the lowering of a test cutoff. And like the two relative differences, the absolute difference tends to change systematically as the prevalence of an outcome changes, though in a more complicated way than the two relative differences. For instant purposes it suffices to note that: (a) when an outcome is uncommon (less than 50 percent for both groups being compared), increases in the outcome will tend to increase absolute differences between rates (at least to the point where one group's rate exceeds 50 percent), while decreases in the outcome will tend to reduce absolute differences; (b) when an outcome is common (greater than 50 percent for both groups being compared), increases in the outcome will tend to reduce absolute differences, while decreases in the outcome will tend to increase absolute differences (at least to the point where one group's rate falls below 50 percent). The reader may note that (b) is implied in (a) since, for example, where rates for an outcome are in a range where decreases in the outcome tend to reduce the absolute difference the rates for the opposite outcome are in a range where the corresponding increase in that outcome tends also to reduce the absolute difference. The illustration in Table 1 happens to involve failure and pass rate ranges

where a decrease in failure rates, with the corresponding increase in pass rates, tends to reduce absolute differences between rates.[4]

Figure 2 illustrates the pattern by which absolute differences change across the entire distribution of test scores using hypothetical test score data according to the same specifications underlying Table 1 and Figure 1.

Figure 2. Absolute Differences between Rates of AG and DG Pass (or Fail) Rates at Various Cutoff Points Defined by AG Fail Rate



The above-described patterns will not be observed in every situation where one examines the favorable or adverse outcome rates of advantaged and disadvantaged groups at different points in time or in settings differentiated other than temporally. Observed patterns are functions of both (a) the described prevalence-related patterns and (b) the comparative size of the differences in the circumstances of the two groups from setting to setting. Understanding the size of those differences, as well what causes them and what can mitigate them, is society's principal, if not sole, interest in examining outcome rates of demographic groups. But one must understand the prevalence-related patterns in order to effectively appraise the size of differences in the circumstances of two groups reflected by their differing outcome rates and determine whether the size of such differences has increased or decreased over time or is otherwise larger in one setting than another.

[4] Nuances of the patterns by which absolute differences tend to change as the prevalence of an outcome changes are discussed in the introductory section of the Scanlan's Rule page of jpscanlan.com. That page also discusses the pattern by which, as the prevalence of an outcome changes, the difference measured by the odds ratio tends to change in the opposite direction of the absolute difference, as shown in Figure 3 (at 7) of the FCSM Paper, and as reflected in the discussion of the paper's Table 1 (at 13). See Figure 5 (slide 27) of the Harvard Applied Statistics Workshop mentioned at page 2 above for an illustration of the effects of lowering a cutoff on all four measures and Figure 6 (slide 38) for a like illustration of the patterns using income data.

In the education context, differences in outcome rates have often been analyzed in terms of relative differences in achieving some outcome or relative differences in failing to achieve the outcome. Invariably, however, those drawing some conclusion or inference based on the comparative size of relative differences in experiencing some outcome have failed to recognize the extent to which the observed pattern is simply a function of the prevalence of the outcome. See the discussion in my "Race and Mortality," *Society* (Jan./Feb. 2000), regarding the way that supporters of affirmative action at elite universities have pointed to the fact that relative differences in graduation rates are lower at more selective than less selective universities, while opponents of affirmative action at elite universities have pointed to the fact that relative differences in failure to graduate are larger at more selective than less selective universities. Those making these points have failed to recognized the extent to which observed patterns are simply a function of the fact that graduation rates tend to be higher at more selective than less selective universities. See also the discussion in my "An Issue of Numbers," *National Law Journal* (Mar. 5, 1990), and "The Perils of Provocative Statistics," *Public Interest* (Winter 1991), regarding the perception that the high proportion that minorities comprise of collegiate athletes disqualified from competing by National Collegiate Athletic Association academic standards has been regarded as a reflection of the fact that the standards were too high, but without recognizing that the lower the standard the greater will tend to be the proportion minorities comprise of those disqualified.

Demographic differences in proficiency rates are sometimes examined in terms of relative differences either as to proficiency or non-proficiency, though invariably without recognizing the role of the prevalence of an outcome. See the Harvard CRP NCLB Study subpage of the Educational Disparities page (EDP) of jpscanlan.com. More often, however, proficiency disparities are studied in terms of absolute differences between rates, but, here too, without recognizing that the way the chosen measure is affected by the rates at issue. Such studies, and their failures of understanding, are discussed on the Educational Disparities page itself and its Disparities by Subject and New York Proficiency Rate Disparities subpages.[5]

The Education Trust's May 2013 study falls into the latter category, as discussed on the Education Trust GC Study subpage. The subpage explains the extent to which the observed patterns of changes in absolute differences are of the kind to be expected in the circumstances regardless of any actual change in the comparative situation of advantaged and disadvantaged groups. Thus, even when the study's findings are broadly correct in the sense of being consistent with those one would reach while employing a measure unaffected by the prevalence of an outcome, such findings are misleading by suggesting that the absolute difference is effectively quantifying the differences in the circumstances of the groups reflected by their outcome rates. Further, as shown in Table 2 of the subpage, more precise rates than the rounded figures discussed in the text of the study would appear to indicate that the study's finding as to increasing disparities between Hispanics and non-Hispanic whites regarding movement into the advanced levels of math is incorrect. And Table 3 of the subpage shows that, while the study

[5] See also my comment on an August 31, 2012 *EdSource Today* post titled "Test scores rise, but achievement gaps persist."

found that increasing disparities between Hispanic and non-Hispanic white rates of reaching the advanced level in math were more pronounced among high income than low income students, a measure of disparity unaffected by the prevalence of an outcome seems to indicate that the disparity in fact decreased among high incomes students.

See also the discussion regarding the subpage's Table 1 and the fact that, in the instance where the study discussed relative differences in rates of failing to reach the basic level, the discussion reflected the mistaken perception that the reduction in disparity measured by the absolute difference would have reduced the relative difference. In fact, for all three situations where the study found reductions in disparity as measured by absolute differences, the relative difference in adverse outcome rates had increased. And, as will necessarily happen when the relative difference in the adverse outcome and the relative difference in the absolute difference have changed in opposite direction, the relative difference in the favorable outcome changed in the same direction as the absolute difference.[6]

The Education Trust's April 2014 study, which examines differing rates of achieving certain favorable academic outcomes among high-achieving students, might appear to raise a somewhat different issue. But, as explained in the Education Trust HA Study subpage, the study failed to recognize that demographic differences in outcome rates among high-achieving students were to be expected – and that the nature of those differences could be fairly predicted – simply on the basis of the differing proportions of each demographic group that fell within the top 25 percent of all students. And that failure of understanding involves the same properties of normal distributions that underlie the patterns by which standard measure of differences between rates tend to change as the prevalence of an outcome changes. The study proceeded on the expectation that, but for some failure of the educational process respecting the fostering of the abilities of high-achieving students from disadvantaged groups, one would expect to observe similar outcome for advantaged and disadvantaged groups within the high-achieving category. For example, the study was premised on the expectation that, absent some failure in the educational system, among high-achieving students, one would expect that, just as 34 percent of high-achieving whites attended highly selective colleges, 34 percent of high-achieving black students would attend such colleges (compared with an actual figure of 19 percent). As explained on the subpage, however, the fact that 35 percent of white students compared with 6 percent of black students fell within the top 25 percent of total students provides a basis to expect that only 16 percent of high-achieving black students would attend elite colleges, a smaller proportion than actually attended such colleges. Table 1 of the subpage indicate that the rates at which high-achieving students from different racial/ethnic or income groups subsequently achieved measures of academic success were generally consistent with expectations based on the proportions of students from the various groups falling within the high-achieving category.

[6] As noted above, the prevalence-related patterns will not always be evident. It is possible for all measure to change in the same direction as the prevalence of an outcome changes, in which case one may infer that there occurred a true change in the comparative circumstances of the groups being examined. But whenever a mentioned relative difference changes in the opposite direction of the absolute difference, the unmentioned relative difference will necessarily have changed in the opposite direction of the mentioned relative difference and the same direction as the absolute difference.

I have not systematically examined studies by Education Trust. But, whatever the precise nature of a study of demographic differences, it cannot provide useful information on such things as whether differences in the circumstances of advantaged and disadvantaged demographic groups have changed over time without consideration of the way the measure employed tends to be affected by the prevalence of an outcome.

In questioning the soundness of Education Trust research, I note that each criticism of that research could be made as well with regard to the methodological approaches of the entities mentioned at the outset as a recipients of a letter of this nature, as well as the National Center for Health Statistics (subject to certain qualifications[7]), the Agency for Healthcare Research and Quality, the Centers for Disease Control and Prevention, and all other research institutions in the United States or abroad, as discussed at pages 26 to 32 of the FCSM Paper. But I urge the Education Trust, in evaluating the soundness of its research to date and in determining how it will conduct research in the future, not to be influenced by the failure of other institutions yet to recognize and address these issues. So far, few statisticians or other researchers at these institutions are aware of the patterns reflected in Figures 1 and 2. Probably, the Department of Education is not yet aware, in any institutional sense, that lowering a test cutoff will tend to increase relative differences in failure rates at the same time that it reduces relative differences in pass rates. But the patterns obviously do exist. And once recognizing them, an institution desiring to advance the understanding of differences in the circumstances of advantaged and disadvantaged groups must consider the implications of those patterns.

In that regard, I call your particular attention to Section B (at pages 12 to 23) of the FCSM Paper. It addresses a common perception in health and healthcare disparities research that different measures may all be valid in their way, even when they yield opposite conclusions about such things as whether disparities have increased or decreased over time, and that a value judgment is involved in choosing among them. There exists a related perception that the complexities of measurement may be addressed simply by presenting both the absolute difference and whichever relative difference the observer happens to be examining. But as explained in that section (and also discussed in many other places, including the Harvard University letter (Section D, at 24 to 28) and the Kansas Law Faculty Workshop Paper (Section B, at 15 to 23)) there exists only one reality as to whether differences in the circumstances of advantaged and disadvantaged have increased or decreased over time and standard measures of differences between outcome rates cannot identify that reality.

Finally, I note that I am familiar with President Haycock's August 27, 2013 letter to the Department of Education on civil rights data collection issues, including its thoughtful comments on the manner in which data on discipline are maintained. While I am unfamiliar with any Education Trust analyses of discipline data, I recognize that at some point the Education Trust may study discipline disparities issues. If Education Trust does so, it should be mindful that the statistical issues addressed in this letter play importantly into interpretations of data on discipline

[7] As discussed at pages 11-12 of the FCSM Paper, the NCHS at least recognized that the two relative differences tend to change in opposite directions as the prevalence of an outcome changes.

disparities. As noted above, the three recent articles discussed at the beginning of the fourth paragraph all relate to the mistaken perception that generally reducing discipline rates will tend to reduce relative racial/ethnic differences in discipline rates. Further, the California Disparities, Maryland Disparities, Los Angeles SWPBS, and Denver Disparities subpages of the Discipline Disparities page of jpscanlan.com discuss the fact that general reductions in discipline rates in each of the referenced jurisdictions were attended by increasing relative racial/ethnic differences in discipline rates. The DOE Equity Report subpage discusses that a Department of Education report itself showed that relative racial differences in discipline rates were greater in school districts without zero tolerance policies than in districts with such policies. And the Preschool Disparities subpage discusses that larger relative racial differences in suspension rates, though smaller relative differences in rates of avoiding suspension, are to be expected in preschool compared with K-12 simply because suspension rates are very low in preschool. Similarly, the Suburban Disparities subpage explains the reasons to expect larger relative racial difference in suspension rates in suburbs than in central cities, given the lower overall suspension rates in suburban schools. I suggest that the Education Trust will find it useful to review all of the Discipline Disparities subpages before analyzing disparities in discipline rates.

Sincerely,

/s/ James P. Scanlan

James P. Scanlan