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

May 14, 2014

Michael L. Askew
Chair of the Board of Directors
Patrick McCarthy
President and Chief Executive Officer
Cindy Guy
Director, Research and Evaluation
The Annie E. Casey Foundation
701 St. Paul Street
Baltimore, MD 12202

Re: Measurement Issues Pertaining to Annie E. Casey Foundation Research

Dear Chair Askew, President McCarthy and Director Guy:

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), the Investigations and Oversight Subcommittee of House Finance Committee (Dec. 4, 2013), and The Education Trust (April 30, 2014).

This letter follow on a similar letter to the Education Trust that addressed with that organization problems with its efforts to appraise demographic differences in achieving certain levels of

[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.

academic achievement. The instant letter is somewhat prompted by an Annie E. Casey Foundation 2014 study titled "Early Reading Proficiency in the United States," on which I recently created a web page, 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 mentioned above).[2]

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.[3] 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 pattern described in the three articles discussed above whereby lowering a test cutoff tends to reduce relative differences in pass rates while increasing relative differences

---

[2] The measurement pages include: Educational Disparities, Measuring Health Disparities, Scanlan's Rule, Mortality and Survival, Immunization Disparities, Disparate Impact, Discipline Disparities, Lending Disparities, Employment Discrimination.

[3] 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).

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 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. [4] 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.

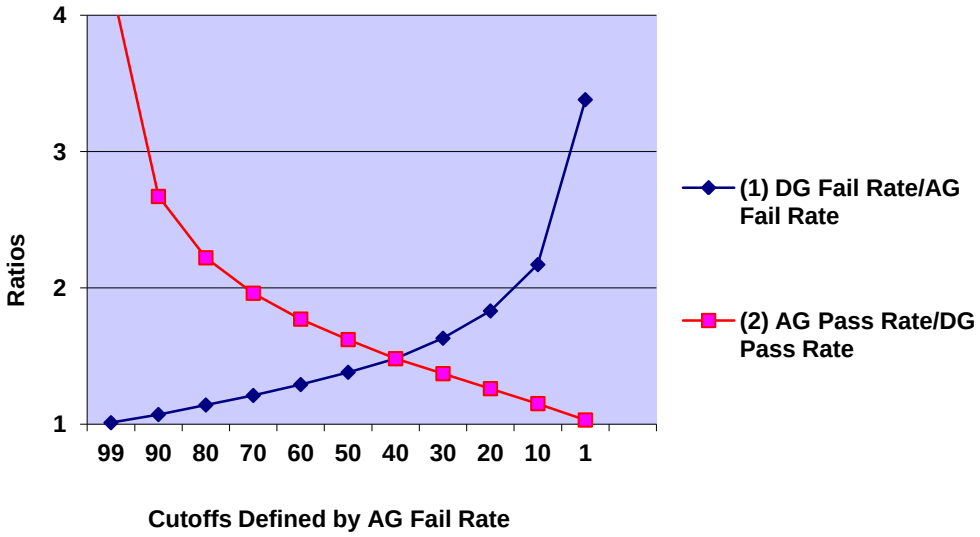
Table 1. Pass and Fail rates of Advantaged Group (AG) and Disadvantaged Group (DG) at Two Cutoffs, with Ratios of (1) DG Fail Rate to AG Fail Rate and (2) AG Pass Rate to DG Pass Rate at Various Cutoff Points Defined by AG Fail Rate, and Absolute Difference between Rates

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.

[4] 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

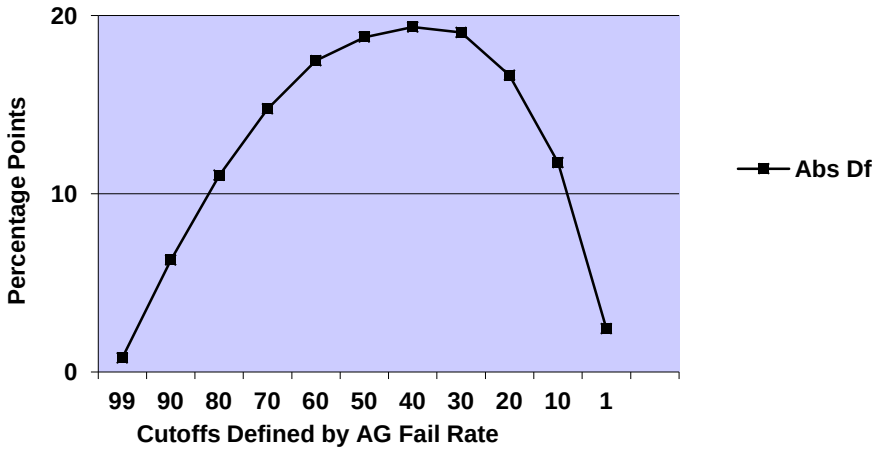


The absolute difference between rates – 17 percentage points with the higher cutoff and 8 percentage points with the lower cutoff in Table 1 – is unaffected by whether one examines the favorable outcome or the corresponding adverse outcome. But for a measure to effectively quantify the strength of the 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 a corresponding increase in pass rates, tends to reduce absolute differences between rates.[5]

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.

[5] 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.

A useful way to appreciate the importance of understanding the prevalence-related forces is to consider the devoting of resources to studying the reasons for (or drawing inferences about and making policy decisions on the basis of such inferences) the fact that, according to some standard measure, the advantaged and disadvantaged benefitted differently from the general improvements in test performance that caused the situation to change from that reflected in the first row of Table 1 to that reflected in the second row. Even leaving aside that different measures would yield different conclusions, there can be no sound interpretation of data on changing patterns without recognizing the patterns that commonly occur even when the strength of the forces causing rates to differ has changed not all.

In the education context, differences in outcome rates have 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," *Societ*y (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.[6]

The Education Trust's April 2014 study titled "Falling Out of the Lead: Following High Achievers Through High School and Beyond" fits into the latter category. The study examined

[6] See also my comment on an August 31, 2012 *EdSource Today* post titled "Test scores rise, but achievement gaps persist." The general failure to understand the way standard measures of proficiency differences change as proficiency rates change will be further addressed in my forthcoming "Race and Mortality Revisited," *Society* (July/Aug. 2014).

changes in racial/ethnic differences in falling below the basic reading level and in reaching the advanced reading level during a period when rates of reaching the basic level and rates of reaching the advance level were generally rising. The study measured racial/ethnic and income differences in terms of absolute differences between rates. As with virtually all other research into difference between outcome rates of advantaged and disadvantaged groups pertaining to education, however, the study did so without recognizing that general increases in achieving favorable educational outcomes (a) will tend reduce absolute differences for outcomes where rates are in ranges at issue for meeting/failing to meet the basic level and (b) will tend to increase absolute differences for outcome where rates are in the ranges at issue for reaching/failing to reach the advanced reading level. The study if further is discussed on the Education Trust GC Study subpage of the Educational Disparities page.

That Education Trust study is also discussed on the recently created Annie E. Casey 2014 Proficiency Disparities Study subpage in the course of discussion of the failure of the Annie E. Casey Foundation study to recognize that standard measures of differences between outcome rates tend to be affected by the prevalence of an outcome. As discussed on the subpage, the recent Annie E. Casey Foundation study is somewhat unclear on what measure it employed, which is a serious problem given that the measures it might be employing tend to change in opposite directions as the prevalence of an outcome changes. See the FCSM Paper's discussion regarding its Tables 3 (at 16-17) and 10 (at 21-22) and regarding Healthy People 2010 (at 25-26) with respect to research that measures relative differences in one outcome while describing relative differences in the opposite outcome. The more serious problem, however, is the broader failure to recognize that standard measures tend to change simply because the prevalence of an outcome changes and hence cannot effectively quantify the difference in the circumstances of advantaged and disadvantaged reflected by their rates of experiencing some outcome.

The subpage includes a table to illustrate the way that absolute differences between rates, which the Annie E. Casey Foundation proficiency study evidently relies upon at least for some purposes, change in the ways described above for the outcomes examined in the Education Trust study. But the subpage also shows the way that the outcome rates examined in the recent Annie E. Casey Foundation proficiency study were in ranges where general improvements tend to increase absolute differences between rates (though, in some states proficiency rates may be approaching levels where further improvements will tend to reduce absolute differences).

The subpage also discusses the 2014 Annie E. Casey Foundation study titled "Race for Results: Building a Path of Opportunity for all Children," which employed a methodology (a) that is in the direction of the approach that I recommend for measuring differences in the circumstances of advantaged and disadvantaged groups reflected by their differing outcome rate and (b) that appears to be aimed at addressing the problem arising from the fact that any value for standard measure reflects different strengths of association at different levels of overall prevalence. I am not yet familiar enough with the approach to form an opinion on its utility. But the thinking about the problems in employing a standard measure to appraise disparities concerning matters with different baseline rates discussed at page 4 of the methodology document – which is the same problem involved with the employing of a standard to appraise the size of the differences in

the circumstances of two groups reflected in the two rows of Table 1 – indicates an important recognition that has implications well beyond that particular study.

In questioning the soundness of Annie E. Casey Foundation research, I note that the criticism of that research could be made as well with regard to the methodological approaches of each entity mentioned at the outset as a recipient 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 Annie E. Casey Foundation, 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 not broadly familiar with the work of the Annie E. Casey Foundation. But the issues I raise concerning appraisal of demographic differences in meeting some level of proficiency pertain to all efforts to appraise demographic differences in educational outcomes. Institutions that study educational disparities issues are increasingly studying racial/ethnic and other difference in discipline rates. Little of that research has been sound. 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

[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.

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, with respect to a matter recently receiving much media attention, that larger relative racial differences in suspension rates, though smaller relative racial differences in rates of avoiding suspension, are to be expected in preschool than 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 Annie E. Casey Foundation 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