NCAASE NATIONAL CENTER ON ASSESSMENT AND ACCOUNTABILITY FOR SPECIAL EDUCATION

Advancing research on growth measures, models, and policies for improved practice

Using Effect Size to Evaluate Achievement Gaps

Joseph Stevens[1], University of Oregon

ABSTRACT

This research brief provides an overview of methods for estimating the size of achievement gaps. Unfortunately many researchers, analysts, and policy-makers use subjective methods (e.g., visual inspection) to evaluate group differences. Another common method is to take the difference between groups in percent proficient, but there are a number of shortcomings to this approach. A more objective and well tested method to compare groups is effect size (ES). I define ES and demonstrate its computation in several situations including between-group comparisons and within-group change over time. I also discuss newer methods that are appropriate for categorical and ordinal data, methods that examine group differences across the whole score distribution, and tools for visualizing achievement gaps.

Perhaps the most widely used method for describing and reporting the size of achievement gaps in education is the difference in percent proficient for two student groups of interest (e.g., students with and without disabilities) on a state achievement test.

There is substantial room for improvement, however, in the way we report and interpret assessment and accountability information about student achievement and achievement gaps between student subgroups (Supplee, 2008). Many researchers, state and local analysts of accountability data, and policymakers interpret group differences by visual inspection and other subjective methods and are unaware of the difficulties and drawbacks in the use of percent proficient (*PP*) as an accountability measure.

An alternative approach to evaluating group differences is effect size (ES), a relatively simple way of quantifying group differences. ES is relatively independent of sample size or the units of the test score scale and provides a common yardstick that allows comparisons across tests, content areas, or states. The purpose of this paper is to discuss the common method of comparing *PP* for two student groups (PP_1-PP_2) and describe some of its problems. Then, I define and discuss ES measures, the benefits of using ES measures, and demonstrate how to calculate ES in several common situations. Finally, I discuss some additional ways to measure and visualize achievement gaps.

Difference between groups in percent proficient (PP_1-PP_2). Because of the NCLB mandate to use percent proficient as the fundamental metric to describe achievement performance, the difference in percent proficient between two student groups is the obvious way to compare performance. However, this procedure has a number of drawbacks (see Ho, 2008). First, because percent proficient is an ordinal scale, differences between groups in percent proficient are likely not comparable at different locations on the score scale. As an analogy, consider the results of a horse race. We could describe the race results in two ways: order of finish or elapsed time to the finish line. Percent proficient is the same kind of information as order of finishing the race (1st, 2nd, etc.). It is an ordinal scale and only conveys rank order. In contrast, a scale score is more like the elapsed time to finish the race. It is an interval scale where numbers represent more information than just rank. So, for example, one doesn't know from the order of finish how far apart horses were at the finish line and equal differences in order of finish (1st and 2nd vs. 5th and 6th) may represent very different elapsed time differences. For the purpose of evaluating achievement performance and achievement gaps, the exact size of a difference

is important and simply knowing a difference in rank order is not sufficient for meaningful interpretation.

Another complication in the use of PP_1-PP_2 to operationalize achievement gaps is that the percentage of students who are deemed proficient is dependent on the location of the proficiency cut-point in the score distribution (Ho, 2008). The *PP* cut-point location may vary from one content area to another, one test to another, one state to another, or from one grade to another. These differences make it difficult to make comparisons using *PP* and complicate the task of correctly interpreting a difference in *PP* between groups.

Finally, several authors have pointed out that PP_1-PP_2 should not be used to evaluate change or progress over time because changes in proficiency rates are unstable and measured with substantial amounts of error (Ho, 2008; Linn, 2003). Fundamentally, PP_1-PP_2 is not an accurate measure of change or a reliable method for measuring the differences between groups (Holland, 2002; Linn, 2007).

Characteristics of Good Metrics

A number of characteristics are desirable for comparing differences in performance between groups or change over time in a single group. An important characteristic of a good metric is that it is objective–comparisons should not be based on visual inspection or subjective interpretation of data. A second desirable characteristic is that the metric should clearly represent the magnitude and direction of the difference of interest. A third desirable quality of a good metric is that it is independent of the measurement scale being used (i.e., the size of the group difference should not be influenced by units of the particular scale being examined). A fourth characteristic of a good metric for measuring group differences is that the estimated size of difference should not be influenced by the sample size of groups being compared.

Effect Size

One metric that possesses many of these characteristics is effect size (ES; see Huberty, 2002 for a historical review). The use of ES has been the primary mechanism that facilitates comparisons of findings across research studies for some time and there are a variety of ES measures tailored to different kinds of data and study designs (see Hedges & Olkin, 1990; Hunter & Schmidt, 2004; Cooper, Hedges, & Valentine, 2009) and ES is a recommended requirement for all empirical studies (Lipsey et al., 2012). We share the goal of Lipsey et al. to:

> …stimulate and guide [researchers] to go a step beyond reporting the statistics that emerge from their analysis…With what is often very minimal additional effort, those statistical representations can be translated into forms that allow their magnitude and practical significance to be more readily understood by practitioners, policymakers, and even other researchers (p. 1).

Perhaps the most common calculation of ES is the standardized mean difference known as Cohen's *d* (note that there are additional variations on the calculation of ES like Hedges' g that are not discussed here). Cohen's *d* is calculated as the difference in means between two groups divided by the pooled standard deviation for the two groups:

$$\text{Cohen's } d = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}}$$

The pooled standard deviation in the denominator is calculated by pooling or combining the standard deviations of the two groups while taking into account the sample size of each group:

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

This results in a description of between group differences expressed on a standard deviation scale. Cohen provided rules of thumb for interpreting ES that described values of about 0.20 as "small," 0.50 as "medium," and 0.80 or greater as "large." However, it should be noted that Cohen recommended that these rules of thumb for small, medium, and large effect sizes are "...for use only when no better basis for estimating the effect size index is available" (Cohen, 1977, p. 25). If one is familiar with a test, its scale, and its standard deviation, an ES may be made more understandable by interpreting the result within a substantive context. For example, a

reading achievement gap ES of 0.50 between groups in Grade 3 may not be nearly as concerning as the same achievement gap at Grade 6, given that there is often much greater annual reading growth in Grade 3 than in Grade 6. In other words, the same ES can have very different meaning depending on the context (i.e., interpreting achievement gap ES is more meaningful within the context of knowledge about the measure and the groups being compared).

Thus, ES measures provide a common yardstick that allows comparisons of group differences across tests, grades, content areas, and states. Using a common yardstick allows more informed consideration of the magnitude of group differences that is not dependent on sample size, test/instrument, content area, or time of measurement and is a more objective, standardized way to describe group differences. Note that there are a variety of other ES measures in addition to Cohen's *d* that are not discussed here (see Cooper, Hedges, & Valentine, 2009; Hunter & Schmidt, 1990; Rosenthal, 1991).

Example of Using ES to Represent Achievement Gaps

Bloom, Hill, Black, and Lipsey (2008) describe several methods for representing performance differences between groups using ES. The calculation they recommend for comparing the performance of two groups is the same as Cohen's *d* presented above with one change; in Bloom et al.'s calculation, the denominator for ES is not the *SD* pooled across the two groups but the standard deviation of *all* participants on the test in that grade/occasion (SD_{All}). There are two advantages of this approach: SD_{All} is now based on a larger, more stable group, and second, SD_{All} now defines a scale (i.e., yardstick) with the same units no matter which groups are being compared. Following is an example of this calculation using data from Stevens et al. (2015) featuring North Carolina students in Grades 3 to 6 with specific learning disabilities (SLD) compared to students without disabilities (SWOD). The table shows means by student group, *N*-size, mean differences, and the standard deviation for all students in each grade (SD_{All}).

Grade	SLD	N_{SLD}	SWOD	N_{SWOD}	$\bar{X}_{SLD} - \bar{X}_{SWOD}$	SD_{All}	*ES*
3	246.42	4,877	251.17	80,405	-4.75	7.495	-0.63
4	252.19	4,397	257.04	75,948	-4.85	8.209	-0.59
5	257.97	4,467	263.06	73,155	-5.09	8.684	-0.59
6	263.75	4,447	267.16	71,170	-3.41	9.178	-0.37

So, if we were interested in estimating the size of the achievement gap between SLD and SWOD students in Grade 3, we would calculate ES = (246.42 – 251.17) / 7.495 = -0.63.

Example of Using ES to Represent the Magnitude of Change over Time

Another method discussed by Bloom, et al. is the use of ES to represent the amount of growth or change across two times. Year-to-year or grade-to-grade "transition" effect size can be estimated by examining the mean difference for a group from one year to the next in ratio to the pooled standard deviation for that group for the two years of interest.

As an example of computing transition ES across grades, results for SWOD are presented in the table below. To measure the differences in achievement from Grade 3 to Grade 4, the calculation would be $(257.04 – 251.17) / [(6.90^2)(73,560) + (7.61^2)(69,453) – 2] = 0.810$.

Grade	Mean	*SD*	$\bar{X}_{Time\,1} - \bar{X}_{Time2}$	*N*	Transition *ES*
3	251.17	6.90		73,560	
4	257.04	7.61	5.87	69,453	0.810
5	263.06	7.92	6.02	66,905	0.776
6	267.16	8.34	4.10	65,141	0.504
7	269.98	9.95	2.82	63,539	0.308

Another commonly used method for comparing performance over two time points (as in examining change from pretest to posttest) is the calculation recommended by Lipsey and Wilson (2001) in which a gain score (posttest mean – pretest mean) is divided by the pretest *SD*. An attractive feature of this calculation is that change over time is expressed as a function of the original units at the initial time point.

Technically, the previous two methods of estimating ES over time may underestimate ES when the same participants are followed over time (i.e., true longitudinal designs). In that situation, participants' performance at one time point is likely correlated with performance at other time points. In this case, there are a variety of other methods for estimating ES. One common calculation designed to take correlation over time into account is to divide the gain score by the following standard deviation denominator that corrects for correlation over time where *Cov* is the covariance across pretest and posttest scores:

$$S_{Time} = \sqrt{S^2_{Time} + S^2_{Time2} - 2Cov(Time1, Time2)}$$

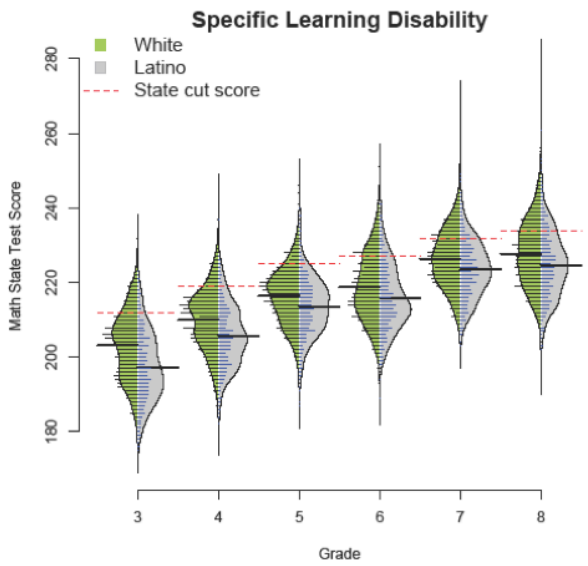
Methods also exist for estimating ES in more complex longitudinal designs and situations where more than two time points are used. These more complex methods are largely beyond the scope of this brief description of ES methods, but one example of these methods is the calculation of a *growth* ES based on the parameters of a longitudinal growth model. This method may be particularly useful when growth is not linear, as is often the case with academic growth. This method uses the estimated growth rate at a particular point in time as the metric of the magnitude of change. In a curvilinear growth model, growth rate (GR) is defined as:

$$GR = S + (2)(Q)(t)$$

where *S* and *Q* are the estimated linear slope and quadratic growth parameters in the growth model and t represents time or measurement occasion. To obtain the ES, the GR is divided by the standard deviation at time t.

Whole Distribution Methods

Another challenge in understanding and representing the size of group differences surrounds the nature of the score distributions for the two groups. Using the ES metrics described above, group differences are evaluated only at one point in each group's score distribution (i.e., the proficiency cut-point or the group mean). When score distributions for each group are normally distributed, ES metrics may represent group differences throughout the distributions fairly well. However, it may still be valuable to examine group differences throughout the score distributions to better understand group differences. For example, consider the graphical display below (called a bean plot due to its appearance). The great advantage of these plots is that they readily allow comparisons of the groups at any level of performance. The figure shows differences between SLD students in Grades 3 to 8 who are Latino (in grey on right) versus White (in green on left) on the Arizona mathematics test.



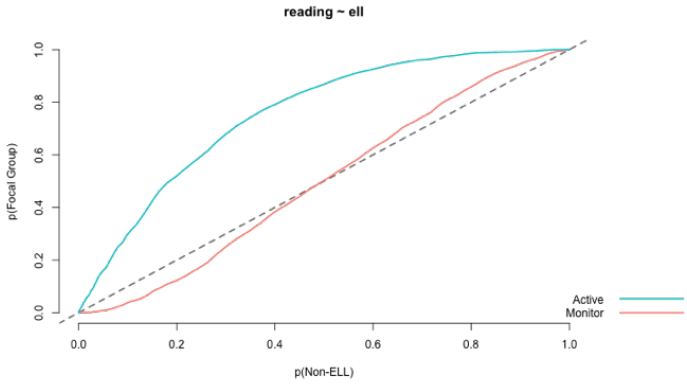
At each grade, the red dotted lines show the state proficiency cut-point (the location of PP_1-PP_2) and the black lines show the mean of each group (the location of ES). As can be seen, the figure quickly conveys information about the whole score distribution for each group and group differences can be examined at any level of performance from the lowest to the highest scorers in each distribution. Note that different conclusions about the achievement gap between these student groups might be drawn depending on whether the comparison was made at the proficiency cut-point, the mean, or at other locations in the score distributions.

Another complication occurs when there are differences in the shape of score distributions for the two groups (i.e., differences in skew and/or kurtosis). In that case, use of PP_1-PP_2 is an even more flawed metric and traditional ES representations like Cohen's *d* may also be inaccurate. As a result, alternative ES measures have been developed when weaknesses are present in the score scale (e.g., ordinal scaling like *PP*) or differences exist in score distributions between groups. When the outcome of interest

is measured on a categorical or ordinal scale, such as the proficiency categories used for accountability reporting, traditional ES methods are not appropriate. Ho and Reardon (2012) have developed methods that estimate the size of group differences across the whole score distribution even using categorical or ordinal data (e.g., area under the curve, the *V* statistic). Their nonparametric ES approach estimates the differences between groups across the score distribution when only "coarse ordinal data" are available (e.g., percent proficient). Ho and Reardon conceptualize *ES as an area* of overlap between two score distributions rather than a single point estimate of the difference between groups. They demonstrate the use of area under the curve (AUC; see example below) and a *V* statistic for this purpose.

A free *R* software package, called *esvi*s (Anderson, Stevens, & Nese, 2017; see http://www.dandersondata.com/page/esvis/) has been developed that computes a variety of ES measures for two or more groups and also produces several visualizations of group differences that can be used to help interpret and understand gaps in achievement. The *esvis* package is designed to visually compare two or more distributions across the entirety of the scale, rather than a single point in the score distribution (i.e., *PP* or group mean). The software package also includes some functions for estimating effect size, including Cohen's *d*, percentage above a cut-point, area under the curve (AUC), and the *V* statistic, which essentially transforms the AUC metric to standard deviation units like a traditional ES (see Ho & Reardon, 2012).

The figure below provides an example of these whole distribution methods examining performance for Arizona students with three different English language learner (ELL) classifications: (a) active ELL student (currently receiving services; green line), (b) monitor (students who previously received services; red line), and (c) non-ELL student (students who never received services; dashed, black line). If the performance of the Active ELL or Monitor ELL groups was the same as the non-ELL group, the colored line for that group would fall directly on the dashed, black Non-ELL line. The bigger the discrepancy between the lines, the larger the achievement gap. The figure shows much larger discrepancies for the Active ELL group than for the Monitor ELL group. In the case of the Active ELL group, the AUC statistic is the entire area between the green and dashed black lines. Notice in this plot there is actually a reversal of the achievement gap difference for monitor students (red line). On the lower end of the scale, Monitor students are actually out-performing non-ELL students (their line is below the dashed, black line), but this effect reverses at the top of the scale. A summary measure like PP_1-PP_2 or Cohen's *d* would not provide this type of more detailed information about group differences.



Conclusions

Subjective methods like visual inspection or descriptive comparisons of *PP* from one group to another are inadequate for evaluating achievement gaps and guiding educational policy and decision-making. It is critically important to apply more sophisticated comparisons than PP_1-PP_2 to characterize achievement growth and/or achievement gaps and apply a measure of group differences that uses a "common yardstick" to allow comparisons across content, tests/instruments, occasions, and locations. In general, use of ES measures is an accepted method that is a substantial improvement over subjective methods. However, choice of ES measure can be complex and should be considered carefully. Specifically, one should consider the purpose of estimating gaps in choosing the best metric or calculation to use. As part of this choice, also consider the nature of the performance scale being used and take the distributional characteristics of the scores into account. When scales are "coarse" as with *PP*, consider using the nonparametric methods described by Ho and Reardon, and when score distributions differ in shape across groups, consider using whole distribution ES methods.

One also needs to be careful in matching ES calculation method to the measurement scale and the research design being used and in particular in distinguishing ES methods for estimating between-group differences from methods appropriate for measuring change over time. Last, because there are many ways to calculate ES, it is very important to clearly report the method/ formula used for calculating ES and be specific about what *SD* is used in the denominator.

Contact Information:
[1]Address correspondence to Joe Stevens, 541-346-2445, stevensj@uoregon.edu, University of Oregon.

References

Anderson, D., Stevens, J. J., & Nese, J. F. T. (2017, April). *Visualizing effect sizes across the full distribution.* Paper presented at the annual meeting of the National Council for Measurement in Education, Washington, DC.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*, 289–328.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd Ed.). NewYork, NY: Russell Sage Foundation.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*, 351-360.

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal ''proficiency'' categories. *Journal of Educational and Behavioral Statistics, 37*, 489-517.

Hedges L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

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

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*, 227-240.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3–13.

Linn, R. L. (2007). *Educational accountability systems.* Paper presented at The CRESST Conference: The Future of Test-Based Educational Accountability.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Lipsey, M. W., & Wilson D. B. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Schulte, A. C., Stevens, J. J. Elliott, S. N., Tindal, J., & Nese, J. F. T. (2016). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test. *Journal of Educational Psychology, 108*, 925-942.

Stevens, J. J., Schulte, A. C., Elliott, S. N., Nese, J. F. T., & Tindal, G. (2015). Mathematics achievement growth of students with and without disabilities on a statewide achievement test. *Journal of School Psychology, 53*, 45-62.

Supplee, L. H. (2008). Introduction to the special section: The application of effect sizes in r esearch on children and families. *Child Development Perspectives, 2*, 164-166.

Funding Source:
This research was funded through the Institute of Education Sciences (IES) (http://ies.ed.gov) through a Cooperative Service Agreement establishing the National Center on Assessment and Accountability for Special Education NCAASE (PR/Award Number R324C110004). The findings and conclusions expressed do not necessarily represent the views or opinions of the U.S. Department of Education.