Using Effect Size Measures to Estimate and Report Achievement Gaps

Joseph Stevens
Daniel Anderson
Joseph Nese
and
Gerald Tindal
University of Oregon

Presented at the annual NCME Conference, San Antonio, TX, April, 2017

Presentation Purpose

- Describe methods:
  - For estimating achievement gaps
  - To more effectively interpret and report gaps including both common and rarely used methods to estimate effect size (ES)
- Demonstrate these methods using:
  - Operational state accountability data from several states in math and reading
  - Achievement differences between several student subgroups
  - Longitudinal academic growth data

(Contact information and acknowledgements at end)

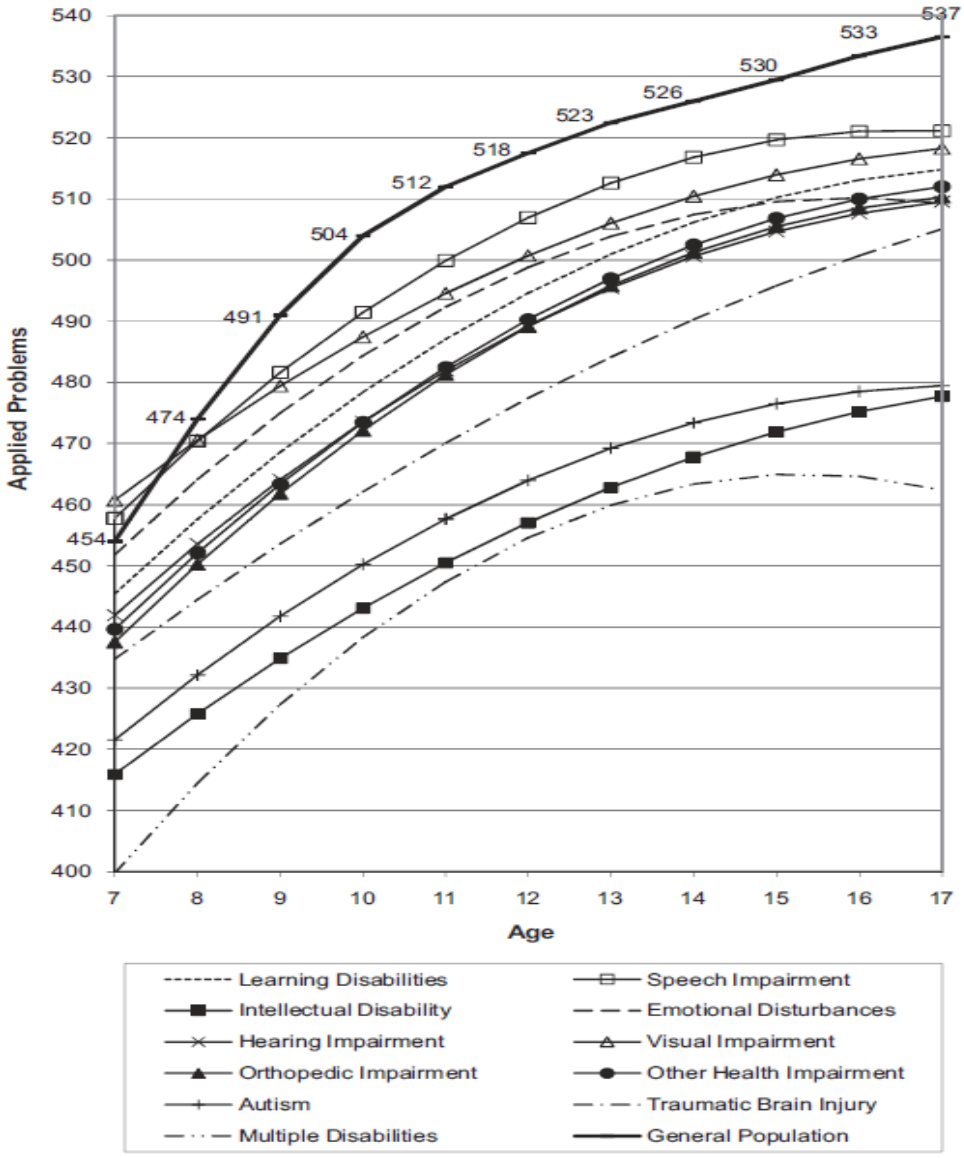
Presentation available at: http://pages.uoregon.edu/stevensj/NCME.pdf

Current Practice

- Substantial room for improvement in the way assessment and accountability information about student achievement and gaps between student subgroups is reported and interpreted
- Many researchers, state and local analysts of accountability data, and policymakers:
  - Interpret group differences by visual inspection and other subjective methods
  - Do not consider where in the distribution the comparison is made, characteristics of the outcome scale, or distribution issues that can substantially impact conclusions about growth and/or gaps

http://pages.uoregon.edu/stevensj/NCME.pdf

Example of a Recently Published Study Using Visual Inspection of Results



Research and Reporting on Student Achievement Gaps

- As part of NCLB, one of the most common methods for reporting achievement gaps is the difference in percent proficient between two groups (P-P)
- Several shortcomings of this approach however:
  - Group differences often evaluated only at one point in distribution (proficiency cutpoint or sample mean)
  - Because P-P ordinal, units may be different at different locations on the scale; thus size of gaps may be due to differences in units rather than performance
  - Methods may require normally distributed data for both groups; thus size of gap may depend on differences in shape of score distributions

http://pages.uoregon.edu/stevensj/NCME.pdf

Characteristics of Good Metrics for Comparing Differences Over Time or Between Groups

- Objectivity - comparisons should not be based on visual inspection or subjective interpretation of data
- Metric should clearly represent the magnitude and direction of the difference of interest
- Scale independence - size of difference should not be influenced by units of the particular scale
- Sample size invariance - size of difference should not be influenced by *N* size of groups or study
- Common scale – difference should be expressed on a scale that is common across comparisons or studies

Empirical Examples Presented Here

- In interests of time, no discussion of details on samples and instruments in this presentation
  - Data presented based on NCAASE work examining four state accountability systems over time (see website)
  - See Schulte et al. and/or Stevens et al. (see references at end) for details on state databases and state assessment instruments behind some of the examples presented here

The Standardized Mean Difference: Cohen's *d*

(note additional variations like Hedges' *g* not discussed here)

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

Where:

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

Examples of ES gap and ES for change over time

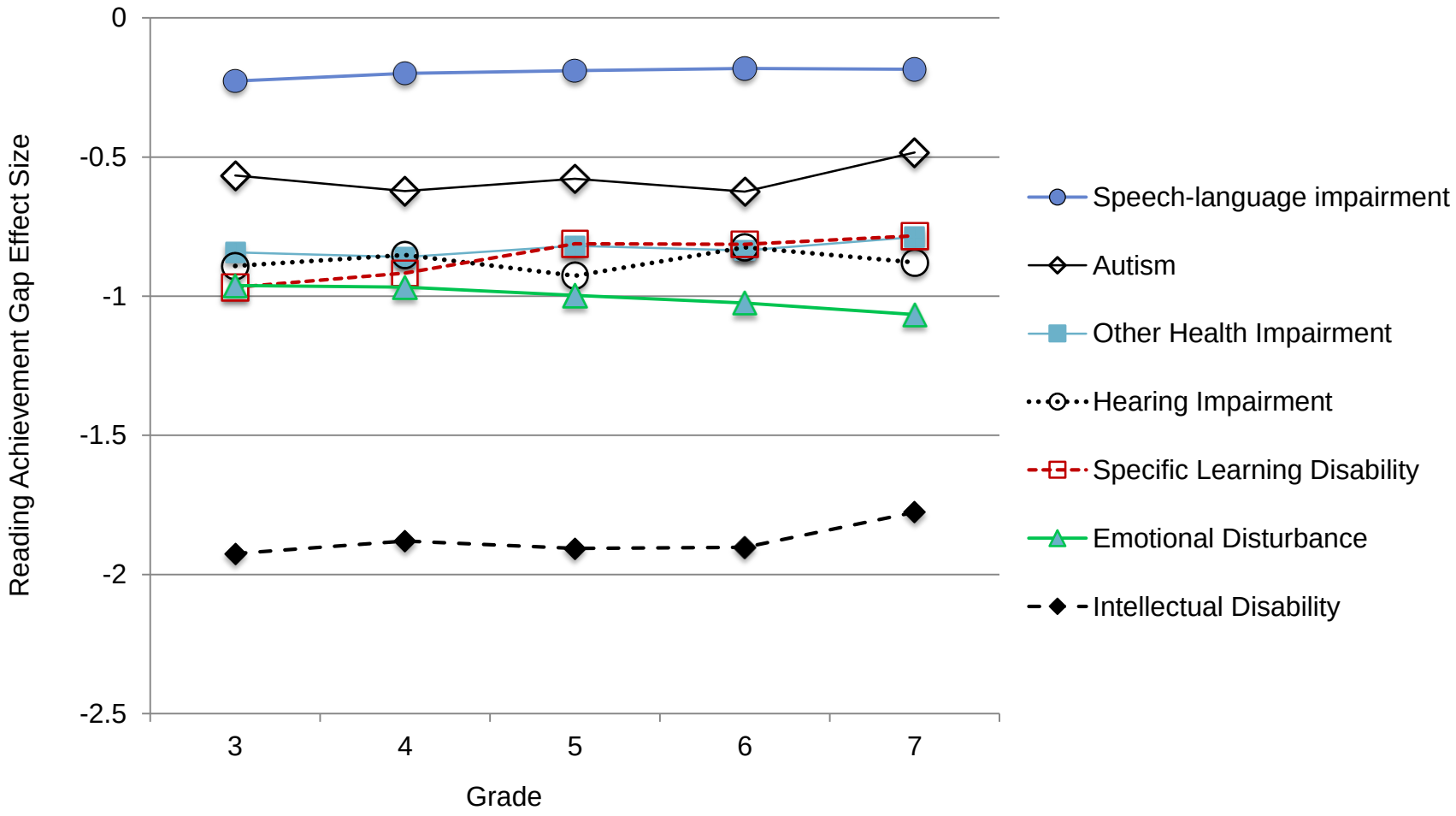
- Bloom, Hill, Black, & Lipsey, 2008:
  - Achievement gap:

    Same as Cohen's d on previous slide except that SD used is the standard deviation of all participants in that grade/occasion (no longer pooling of just the two groups of interest but estimate of population value of the outcome)

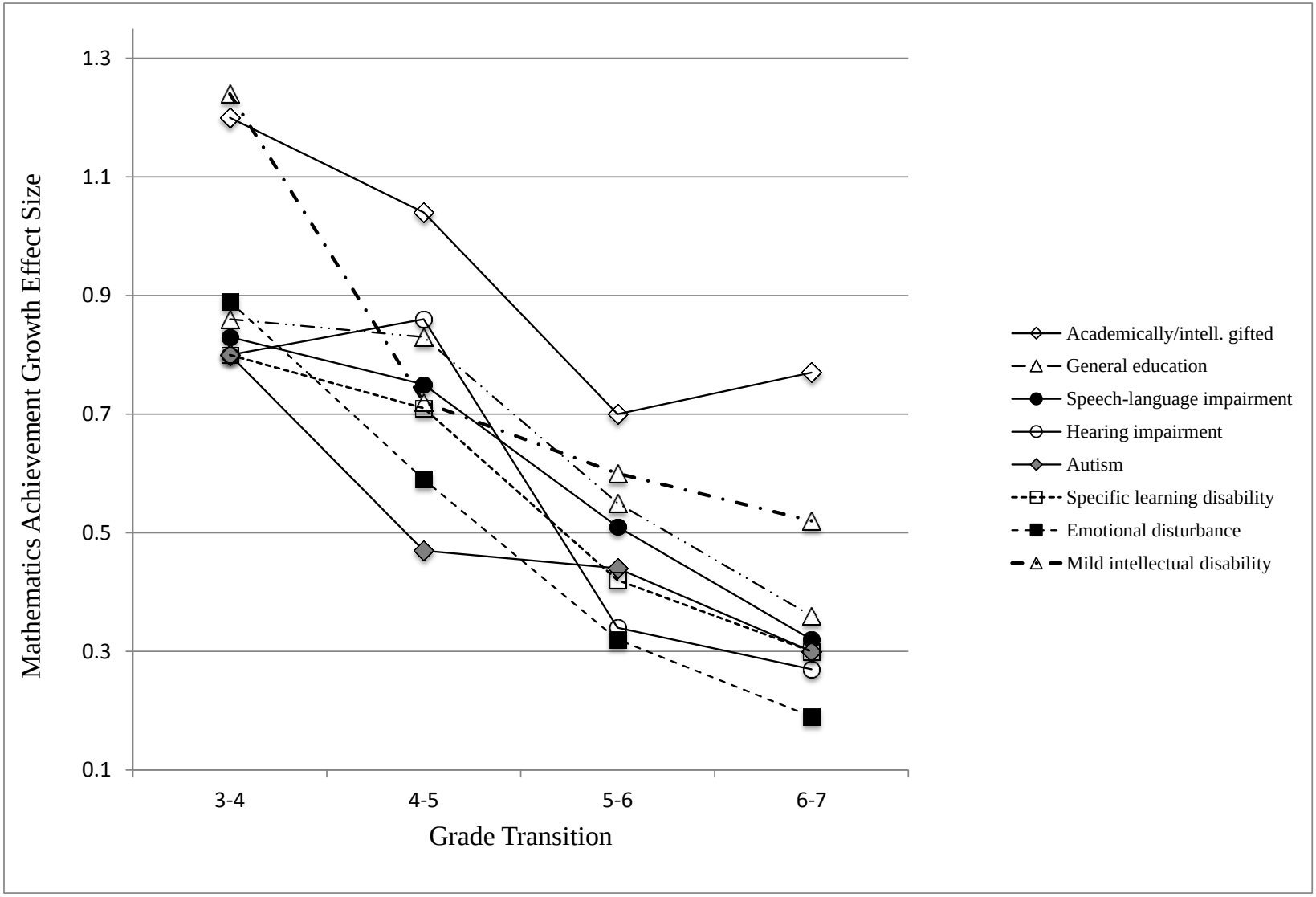
  - Change over time:

    Year-to-year "transition" ESs by examining the mean difference in a group from one year to the next in ratio to the pooled standard deviation for the two years for the group of interest
- Illustrated on next slides

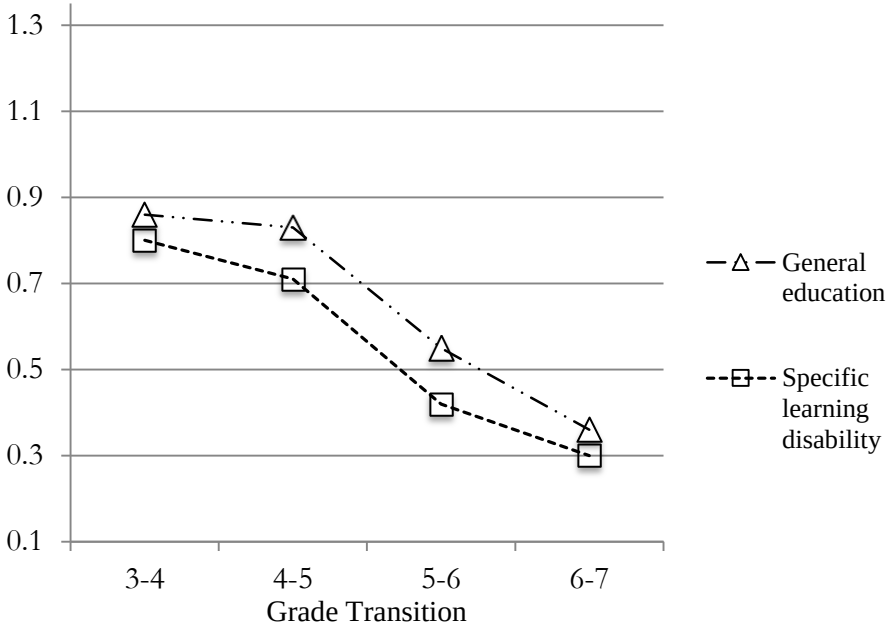
Reading Achievement Gap ES between SWD Groups (compared to regular education students)



ES for Change Over Time by Group



ES for Change Over Time by Group: Calculations

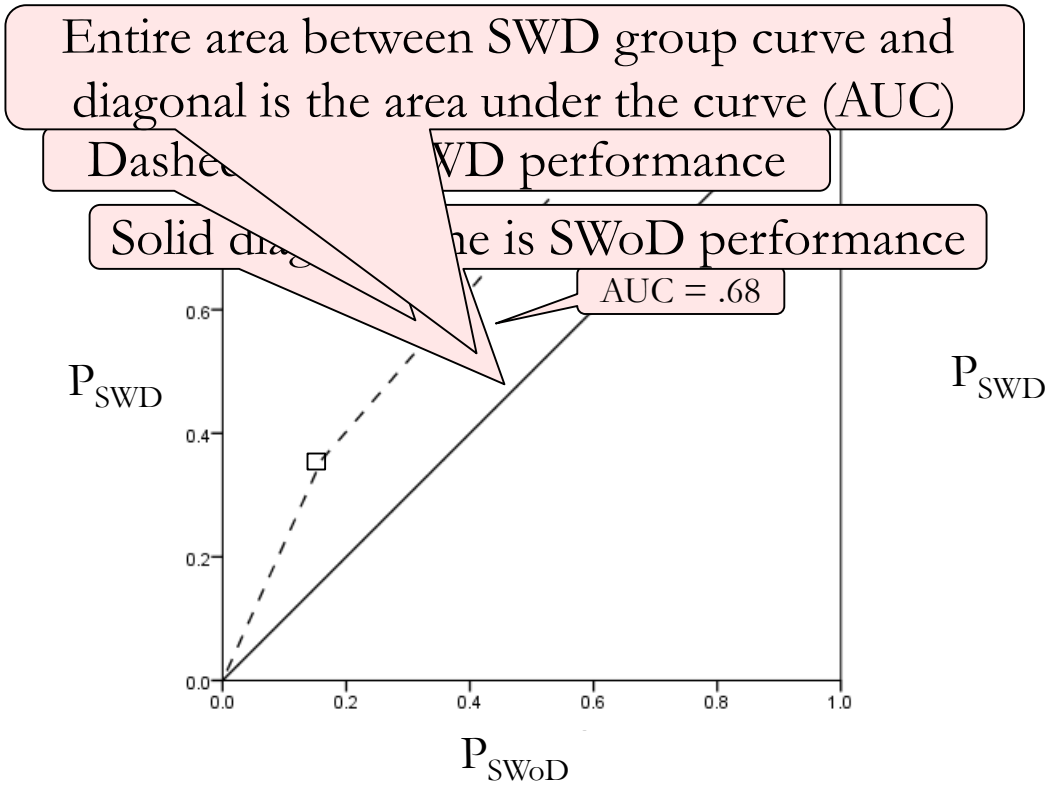


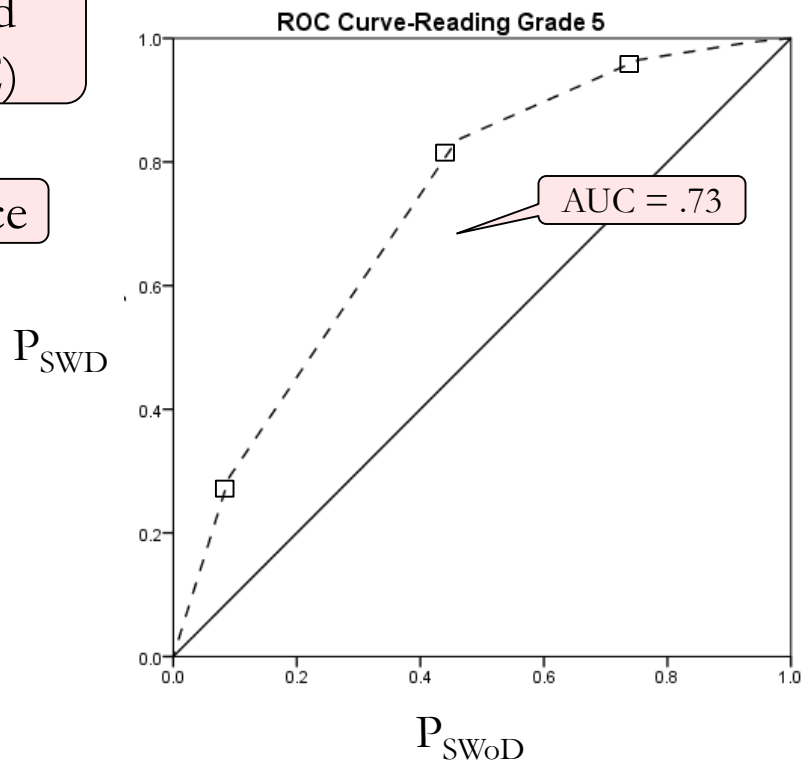
	General Education					Specific Learning Disabilities			
Grade	Mean	SD	$\overline{X}_{t+1} - \overline{X}_t$	Transition ES		Mean	SD	$\overline{X}_{t+1} - \overline{X}_t$	Transition ES
3	251.17	6.90				247.03	6.80		
4	257.04	7.61	5.87	0.810		252.23	7.08	5.20	0.750
5	263.06	7.92	6.02	0.776		257.22	8.07	4.99	0.657
6	267.16	8.34	4.10	0.504		260.28	8.44	3.06	0.371
7	269.98	9.95	2.82	0.308		262.44	9.11	2.16	0.246
		Mean:	**4.70**	**0.599**				**3.85**	**0.506**

Achievement Gaps as Areas Between Score Distributions

- As noted above, a limitation of traditional measures is they only compare groups at the mean or at the proficiency cutpoint, possibly overlooking important group differences lower or higher on the score scale
- Alternative ES measures use whole score distribution and some also accommodate ordinal scales (e.g., proficiency categories; see Ho & Reardon, 2012):
  - Area under the curve (AUC) in Receiver Operating Curve (ROC) analysis
  - V statistic, $V = \sqrt{2}\ (\Phi^{-1})(P_a > P_b)$
- Because of time constraints, we only report a few examples of these analyses

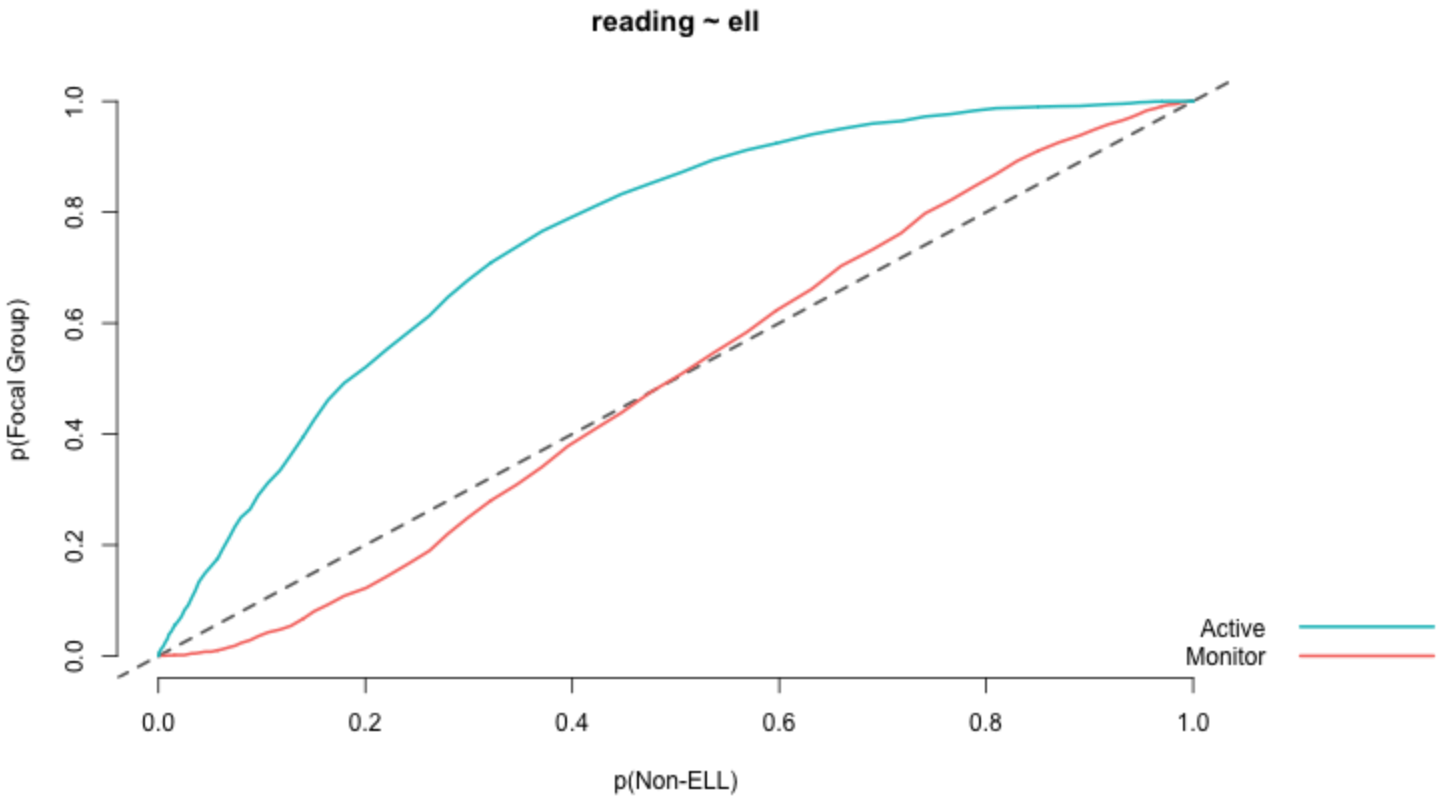
Achievement Gap for SWD vs. SWoD in Oregon Reading in Grade 3 (on left) and Grade 5 (on right)





Whole Distribution Comparisons of Achievement for ELL and Former ELL (monitor) students

(Non-ELL students on Diagonal)



esvis R package (Anderson, see appendix)

ES for Growth

- In addition to Bloom's "transition" ES described earlier, can estimate model-based growth ES using HLM or SEM methods
- There are several growth effect size calculations and variations discussed in literature:

$$\pi_{10} / \mathrm{SD}_{\mathrm{outcome}}$$

  - Note that choice of SD depends on purpose, SD at wave 1 (baseline) is one common choice, but note that SD's may vary over occasions; SD at last occasion or SD pooled over occasions also can be used
  - Also note that ES formulas for estimating power in SEM and HLM, e.g., $\pi_{10} / (\tau_{11}^{1/2})$, are not appropriate as a measure of ES (see Feingold, 2009)

ES for Growth

- Growth rate same at any occasion in a linear model
- In a quadratic model, growth rate differs depending on centering of time (e.g., initial, average, ending) or analytic interest in a particular occasion
- Quadratic growth rate (QGR):

$$\text{QGR} = \pi_{10} + 2(\pi_{20})(\text{time})$$

- Quadratic ES:

$$\text{QGR} / \text{SD}_{\text{outcome}}$$

- Also note differences between unconditional and conditional ES in growth models (latter in contrast to traditional meta-analysis)

Linear Growth Models

Model	π_{10}	Growth Rate	ES Grade 3 SD	ES Grade 5 SD	
Uncond. linear	4.762	4.762	0.690	0.602	*
Cond. linear	5.005	5.005	0.725	0.632	

* Note that calculation of the "power" formula for growth ES results in an overestimate:

$\pi_{10} / \tau_{11}^{1/2} = 4.762/\sqrt{1.272} = 4.222$

Quadratic Growth Models

Model	π_{10}	π_{20}	Growth Rate Grade 3	Growth Rate Grade 5	Growth Rate Grade 7	ES-Grade 3 SD Grade 3	ES-Grade 3 SD Grade 5	ES-Grade 3 SD Grade 7	ES-Grade 5 SD Grade 3	ES-Grade 5 SD Grade 5	ES-Grade 5 SD Grade 7
Uncond. Quadratic	6.925	-0.546	6.924	4.741	2.558	1.004	0.687	0.371	0.875	0.599	0.323
Cond. Quadratic	7.051	-0.518	7.051	4.978	2.905	1.022	0.721	0.421	0.891	0.629	0.367

Conclusions

- Subjective methods like visual inspection to be avoided
- Critically important to apply more sophisticated comparisons than P-P to characterize achievement growth and/or gaps
- Take purpose of estimating gaps or characterizing growth into account in choosing the best metric or calculation
- Consider performance at multiple points in distribution
- Consider scale and distributional characteristics
- Clearly report method/formula for calculating ES and be specific about what SD is used in denominator

Contact Information:

Joseph Stevens, Ph.D.
College of Education
5267 University of Oregon
Eugene, OR 97403
(541) 346-2445
stevensj@uoregon.edu

Presentation available at: http://pages.uoregon.edu/stevensj/NCME.pdf

This research was funded in part by a Cooperative Service Agreement from the Institute of Education Sciences (IES) 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.

Selected References

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.

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods, 14*(1), 43–53.

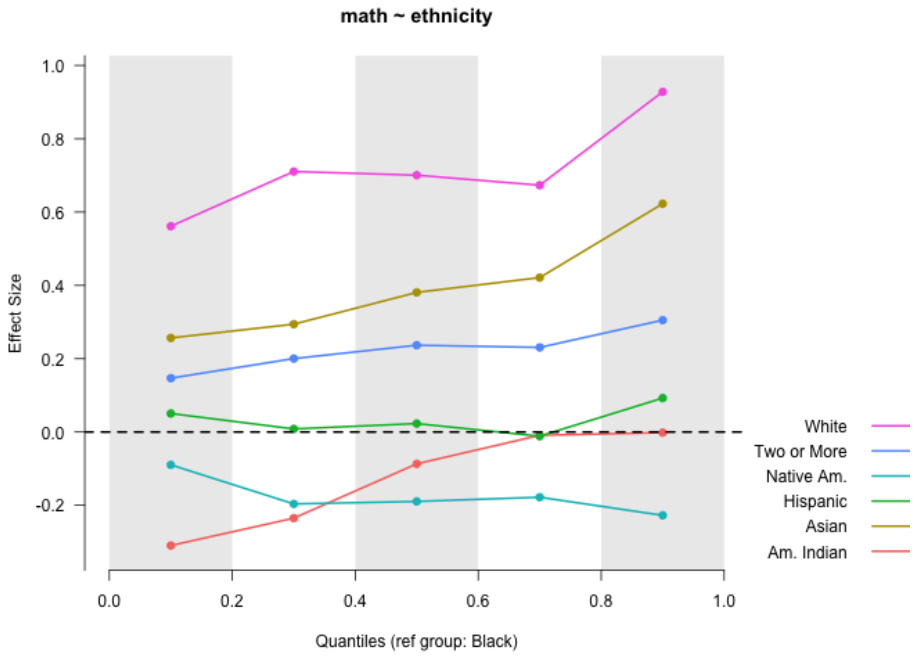
Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "Proficiency" categories. *Journal of Educational and Behavioral Statistics, 37*, 489-517.

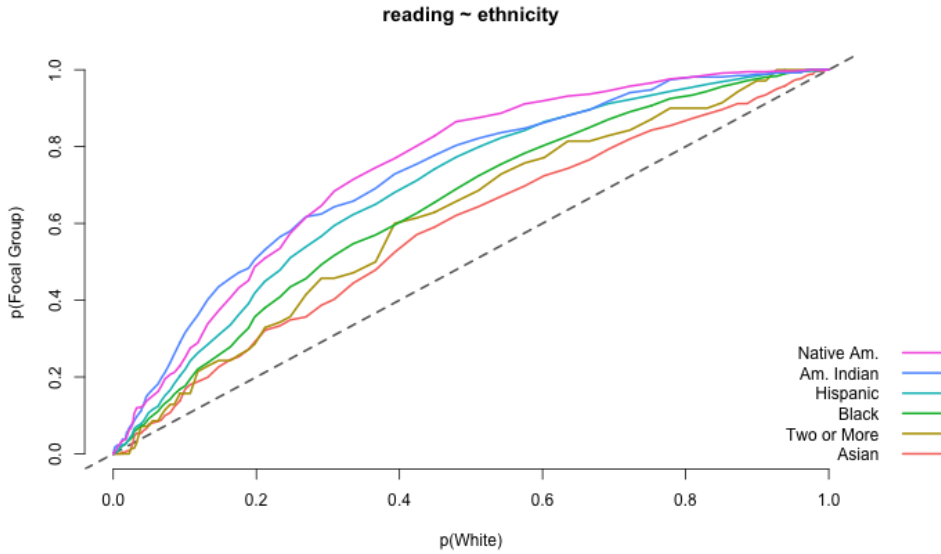
Schulte, A. C., Stevens, J. J. Elliott, S. N., Tindal, J., & Nese, J. F. T. (2016). Achievement gaps for students with disabilities: Stable, widening, or narrowing on a state-wide reading comprehension test. *Journal of Educational Psychology, 108*(7), 925-942.

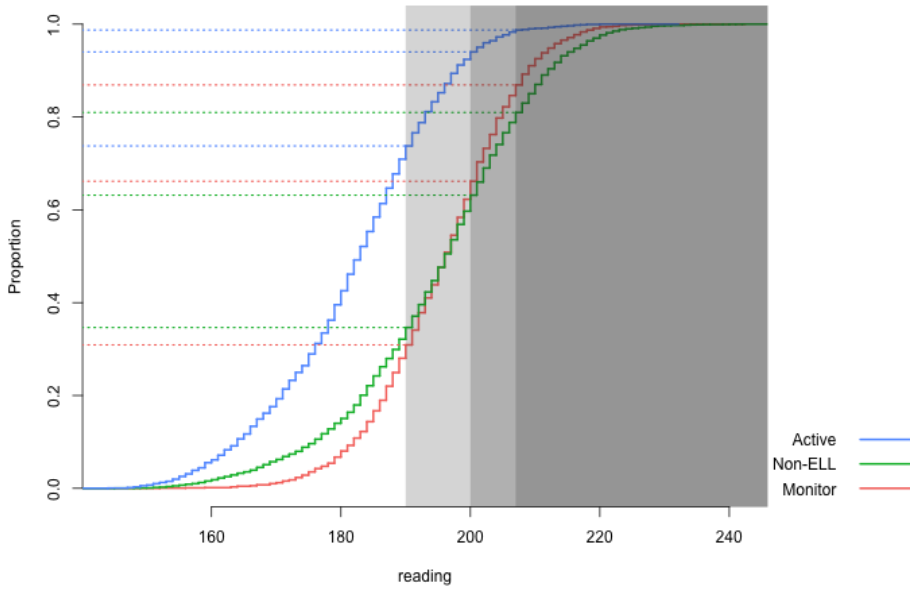
Stevens, J. J., Schulte, A. C., Elliott, S. N., Nese, J. F. T., & Tindal, G. (2015). Mathematics achievement growth of students with and without disabilities on a statewide achievement test. *Journal of School Psychology, 53*, 45-62.

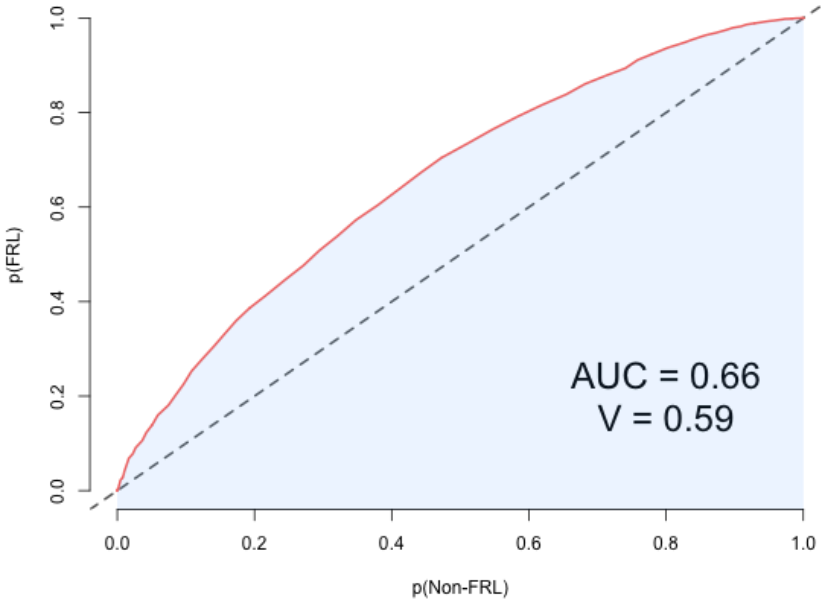
Appendix

Group Comparisons and ES Measures Available From esvis Package









esvis R package (Anderson)

- Plots
  - PP, ECDF, Quantile-binned ES
- Effect sizes
  - Cohen's *d* & Hedges' *g*
  - PAC and TPAC
  - AUC and *V*
- Still under active development
  - Release to CRAN planned for summer

- Install from github

```
install.packages("devtools")
devtools::install_github(
"DJAnderson07/esvis")
```

- Consistent syntax

```
pp_plot(outcome ~ group,
    dataset)
ecdf_plot(outcome ~ group,
    dataset)
binned_plot(outcome ~ group,
    dataset)
coh_d(outcome ~ group,
    dataset)
```