Title: Translating Standardized Education Program Effects into More Interpretable Metrics

Authors: Matthew Baird and John Pane, RAND Corporation

Background: It is common practice in education research to report the effects of programs, policies, or interventions as standardized effect sizes. While this practice eases comparisons across studies, educators and policy makers often struggle with how large or meaningful effects are when measured on this scale. There is a resulting demand for translating standardized effects into more easily interpreted measures. One common example is to estimate the number of years (or days) of learning that would be necessary to induce a similar effect (years of learning option). While this kind of translation can offer clear benefits in ease of interpretation, there are several potential drawbacks, and it is not immediately clear that this conversion should be preferred among various alternatives. Other options for translating to more interpretable metrics include benchmarking the results against gaps between demographic groups or effect sizes measured in other studies (benchmarking), percentile change (percentile), and calculating the likelihood of scoring above a reference value, such as scoring proficient (thresholds). A gap exists in the literature for a thorough documentation of the strengths and weaknesses of the different translation options.

Research Questions: What are the strengths and weaknesses of the four options identified for translating standardized education program effects into more interpretable metrics? Overall, which metrics should be preferred?

Setting: In order to explore the translation options, we take the example of a multi-school educational evaluation, where students from the intervention are exposed to personalized learning education systems. We use data from students that participate in the Northwest Evaluation Association Measures of Academic Progress (MAP). Control students are provided by NWEA to be similar to the treated students on observables. We use this intervention, discussed and evaluated in Pane et al. (2015, 2017), to explore the capabilities and limitations of various translation options.

Population: Our data comes from students across the United States and across all grades in 2012-2015 that participate in MAP Math or Reading evaluations. The following table shows the number of students and schools. There are approximately 85 treated schools, with control students drawn from over 6,000 schools, and over 22 thousand treated students.

<table>
<thead>
<tr>
<th></th>
<th>Math</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
<td>Control</td>
</tr>
<tr>
<td>Number of Students</td>
<td>22,665</td>
<td>378,147</td>
</tr>
<tr>
<td>Number of Schools</td>
<td>86</td>
<td>6,817</td>
</tr>
</tbody>
</table>

Intervention: We use for our example an evaluation of personalized learning in grades K-11. This intervention is discussed in Pane et al. (2015, 2017). The focus of the proposed presentation however is on the options for translating the estimated treatment effects by grade and subject into more interpretable forms.
**Research Design:** We use a calipered coarsened exact matching estimator which regresses standardized posttest scores on pretest scores, treatment status, and control variables, using the matching weights in the regression. We evaluate the effects by grade and subject. We then take four options for translating the effects and make the translation. Those options are benchmarking against demographic groups or effect sizes in the literature, translating to years of learning, translating to percentile change, and estimating the change in the likelihood of scoring above an effectiveness threshold. For the years of learning translation, we test several candidates for scaling options (the measure of typical growth in the grade/subject). For the percentile change, we test two options, the standard normal distribution and an empirical cumulative density function. We then discuss seven desirable properties for any translation option. These are (1) easily interpreted, (2) not easily manipulated post-estimation, (3) smaller confidence intervals, (4) bounded results, guards against insensible results, (5) analysis matching interpretation, (6) can deal with high variance/low average growth tests, and (7) uses all information and is not sensitive to the distribution of scores near thresholds. Examining our findings for our specific case as well as discussing lessons learned in general from the empirical framework of the translations, we rank each of the four translation options within each of the 7 desirable characteristics.

**Data Collection:** NWEA collects data on student performance on the MAP assessments, and provides a sample of these to us.

**Findings:** We find that the rankings in terms of which translation choice is preferred varies across the desirable properties. However, on average the percentile conversion is the most preferred among the four and is ranked first or tied for first for 5 of the 7 properties. On the other hand, the years of learning conversion is ranked last on average, and is ranked last for 5 of the 7 properties.

**Conclusions:** We find that, although the years of learning conversion is very common, it should be avoided by researchers and practitioners because of several severe limitations. When a translation of standardized effect sizes is desired to make results from educational interventions more interpretable, we suggest primarily focusing on converting to percentile growth. Where individuals have preferences over specific characteristics of a translation option, our findings will help guide them to which option is preferred given those preferences. However, the years of learning translation option is dominated by at least one other option for virtually every combination of preferences.

**References:**
