The Relative Performance Index: Neutralizing Simpson’s Paradox

Ernest Davenport
Kyle Nickodem

Department of Educational Psychology
University of Minnesota

February 17, 2017
NAEP – The National Assessment of Educational Progress

The Nation's Report Card is the only nationally representative and continuing assessment of what America's students can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, U.S. history, civics, geography, and the arts (evaluate educational systems, document educational trends)

NAEP is often used as the “gold standard” to evaluate states with respect to Every Student Succeeds (ESS) legislation
California students among worst performers on national assessment of reading and math

Minnesota 4th-graders best in the country in math tests

Scores of Minnesota 8th-graders on the NAEP are fifth-highest in U.S.

Alabama 8th-graders rank 50th in math; see how we compare with other states in reading, science

Surprise: Florida and Texas Excel in Math and Reading Scores

States with the Smartest Kids

Every other year, 4th and 8th graders are tested in reading and math, and
The Problem with Vermont (or Is It Maine?)

Vermont Eighth Graders Rank High Globally in Math, Science

Posted by Meredith King on November 4, 2013

“Vermont eighth graders placed 7th globally in math when ranked against the 49 states and 47 countries and jurisdictions.”

“A recent U.S. study shows that eighth graders in three states, including Vermont, outperformed most students across the world in math and science.”

http://learn.uvm.edu/blog-education/vermont-eighth-graders-rank-high-globally-in-math-science
## 2011 NAEP Results

<table>
<thead>
<tr>
<th></th>
<th>Vermont</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>%</td>
<td>Mean</td>
</tr>
<tr>
<td>White</td>
<td>295</td>
<td>293</td>
</tr>
<tr>
<td>Overall</td>
<td>294</td>
<td>284</td>
</tr>
</tbody>
</table>

---

Department of Educational Psychology
The Problem with Vermont (or Is It Maine?)

“Results from the 2013 National Assessment of Educational Progress, or NAEP, also known as the Nation's Report Card, were released Tuesday and Vermont students scored within the top seven of all states and the District of Columbia in all four categories (Reading / Math in Grades 4 / 8).

http://www.pressherald.com/2013/11/07/maine_students_score_high_on_national_reading__math_tests_/
## 2013 NAEP Results

<table>
<thead>
<tr>
<th></th>
<th>Vermont</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>296</td>
<td>293</td>
</tr>
<tr>
<td>Overall</td>
<td>295</td>
<td>284</td>
</tr>
</tbody>
</table>

### Math

<table>
<thead>
<tr>
<th></th>
<th>%</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>0.92</td>
<td>0.53</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 8th Grade

<table>
<thead>
<tr>
<th></th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>0.92</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
</tr>
</tbody>
</table>
A Problem Hiding in Plain Sight

Simpson’s Paradox

• Simpson’s Paradox is a phenomenon where one population can have higher mean values across all (or most) of its subpopulations yet another population has a higher overall mean.

• This typically arises when unequal sized subpopulations are aggregated.
## 2013 NAEP Fourth Grade Reading Assessment
Who Does A Better Job of Educating Their Students?

<table>
<thead>
<tr>
<th>Subpopulation</th>
<th>Georgia Mean</th>
<th>Georgia SE</th>
<th>Georgia %</th>
<th>Pennsylvania Mean</th>
<th>Pennsylvania SE</th>
<th>Pennsylvania %</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>233</td>
<td>1.460</td>
<td>44</td>
<td>233</td>
<td>1.234</td>
<td>69</td>
</tr>
<tr>
<td>Black</td>
<td>209</td>
<td>1.750</td>
<td>34</td>
<td>208</td>
<td>3.022</td>
<td>16</td>
</tr>
<tr>
<td>Hispanic</td>
<td>213</td>
<td>2.034</td>
<td>15</td>
<td>208</td>
<td>3.649</td>
<td>8</td>
</tr>
<tr>
<td>Asian</td>
<td>245</td>
<td>3.519</td>
<td>4</td>
<td>236</td>
<td>3.422</td>
<td>4</td>
</tr>
<tr>
<td>Multiracial</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>Overall</td>
<td>222</td>
<td>1.101</td>
<td>___</td>
<td>226</td>
<td>1.300</td>
<td>___</td>
</tr>
</tbody>
</table>
Prevalence / Problem

• At least since Yule 1903

• First NAEP 1969 - Report the problem 1973

• Terwilliger and Schield (2004) using the 2002 NAEP 8th grade reading data found 52 statistically significant differences reversed when accounting for Simpson’s Paradox with ethnicity (about 10% of the statistically significant comparisons)
Reporting Subpopulations

ESSA requires reporting scores by:

- Race/ethnicity
- Gender
- Socioeconomic Status
- Disability
- English Language Learners
- Homeless Status
- Military Dependent Status
- Foster Care Status

5 * 
2 * 
3 * 
2 * 
2 * 
2 * 
2 *

\[ 5 \times 2 \times 3 \times 2 \times 2 \times 2 \times 2 = 260 \]
One Index Mania

Desire for one index

• Media
• Policy people
• Parents

But, is it an accurate representation?
Balanced Index

- Tukey, Abelson, Coffman, Gilbert, Jones, & Mosteller (1973)
- Wainer (1986a)

\[ BI_s = \sum W_{iN} \bar{X}_{is} \]

\( \bar{X}_{is} \) is the mean for subpopulation \( i \) in entity \( s \).

\( W_{iN} \) is the proportional representation of subpopulation \( i \) that remains constant across entities.
Balanced Index Example

\[ BI_{GA} = (0.51 \times 233 + 0.15 \times 209 + 0.25 \times 213 + 0.05 \times 245 + 0.03 \times 223) = 224.6 \]

\[ BI_{PA} = (0.51 \times 233 + 0.15 \times 208 + 0.25 \times 208 + 0.05 \times 236 + 0.03 \times 222) = 222.7 \]

Which state is higher now?
Chingos’s (2015) predicts the assessment score based solely on the examinee’s demographics and determines whether the student scores at, above, or below the predicted value (returns a residual).
Problems Lead to a Need for a Better Solution

- $B_l_s$ applies a distribution of observations that is not representative of the population.

- $B_l$ leads to larger standard errors because its weights are not optimal.

- Chingos’ (2015) complex, interactions increase with sources, residuals.

- Subpopulation means are non-parsimonious as variables are crossed.
Score$_s$ is an intermediate value for a given entity

\[
Score_s = \sum W_{is} \left( X_{is} - U_i \right)
\]

$W_{is}$ is the proportional representation of subpopulation $i$ for entity $s$. $X_{is}$ is the mean for subpopulation $i$ for entity $s$. $U_i$ is the standard for subpopulation $i$. 
Relative Performance Index

\[ RPI_s = \frac{S_X}{S_{Score}} \ (Score_s - \overline{Score}) + \overline{X} \]

We transform \( Score_s \) to a standard score
Then transforms this standard score to the scale of the original data.
Standard Errors

The standard error for Scores is

$$SE_{(SCORES)} = \sqrt{\sum (W_{is}^2 \times SE_{X_{is}}^2)}$$

$SE_{X_{is}}^2$ is the standard error for the mean of subpopulation $i$ in entity $s$. 
The standard error for RPI is

\[ SE_{(RPI_s)} = \sqrt{\frac{S^2_X}{S^2_{SCORE}}} \times \sum (W^2_{is} \times SE^2_{Xis}) \]

We are not assuming homogeneous variances for the subpopulations.
Relative Performance Index Example

\[
Score_{GA} = 0.44(233-231) + 0.34(209-205) + 0.15(213-207) + 0.04(245-237) + 0.03(223-225) = 3.40
\]

\[
RPI_{GA} = \frac{6.67}{5.61} \left( 3.40 - (-1.56) \right) + 221.31 = 227.21
\]

\[
Score_{PA} = 0.69(233-231) + 0.16(208-205) + 0.08(208-207) + 0.04(236-237) + 0.02(222-225) = 1.84
\]

\[
RPI_{PA} = \frac{6.67}{5.61} \left( 1.84 - (-1.56) \right) + 221.31 = 225.35
\]
Hypothesis Test Favoring Georgia but not Significant

\[ SE_{(RPI_{GA})} = \sqrt{\left(\frac{6.67}{5.61}\right)^2 \times (0.44 \times 1.46)^2 + (0.34 \times 1.75)^2 + (0.15 \times 2.03)^2 + (0.04 \times 3.48)^2 + (0.03 \times 3.81)^2} \]

\[ SE_{(RPI_{PA})} = \sqrt{\left(\frac{6.67}{5.61}\right)^2 \times (0.69 \times 1.23)^2 + (0.16 \times 3.02)^2 + (0.08 \times 3.65)^2 + (0.04 \times 3.42)^2 + (0.02 \times 4.01)^2} \]

\[ t = \frac{227.21 - 225.35}{\sqrt{1.123^2 + 1.229^2}} = 1.117 \]
Results Table

- 2013 NAEP 4th grade reading assessment
- For every state and jurisdiction:
  - Calculated BI and RPI
  - Determined BI and RPI ranks
  - Compared BI and RPI to original NAEP values
- Correlation with % White
## GA and PA

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
<th>Rank</th>
<th>Rank</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NAEP</td>
<td>BI</td>
<td>RPI</td>
<td>NAEP</td>
</tr>
<tr>
<td>GA</td>
<td>222</td>
<td></td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>PA</td>
<td>226</td>
<td></td>
<td></td>
<td>12</td>
</tr>
</tbody>
</table>
Correlation with % White

<table>
<thead>
<tr>
<th>% White and NAEP state means</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>% White and BI</td>
<td>-0.24</td>
</tr>
<tr>
<td>% White and RPI</td>
<td>-0.13</td>
</tr>
</tbody>
</table>
The Problem with Vermont (and with DC)

<table>
<thead>
<tr>
<th>Percent Tested</th>
<th>WHT</th>
<th>BLK</th>
<th>HIS</th>
<th>ASI</th>
<th>MUL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>10</td>
<td>73</td>
<td>14</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>VT</td>
<td>91</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Nat’l</td>
<td>51</td>
<td>15</td>
<td>25</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>Rank</th>
<th>Rank</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAEP</td>
<td>BI</td>
<td>RPI</td>
<td>NAEP-</td>
</tr>
<tr>
<td>DC</td>
<td>206</td>
<td>51</td>
<td>BI</td>
</tr>
<tr>
<td>VT</td>
<td>228</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>
RPI Flexibility

• Aligns with ESSA emphasis on reporting subgroup performance

• Can be used:
  - With a variety of tests
  - With any demographic variable
  - With comparison values from various sources
  - At national, state, or local levels
Conclusion

The Relative Performance Index

1. Produces a single summary value for each entity (e.g. state)
2. Neutralizes the effect of Simpson’s Paradox
3. Gives voice to subpopulation results