Impact of Methods of Scoring Omitted Responses on Achievement Gaps

Dr. Nathaniel J. S. Brown (nathaniel.js.brown@bc.edu)
Educational Research, Evaluation, and Measurement, Boston College

Dr. Dubravka Svetina & Shenghai Dai
Inquiry Methodology Program, Indiana University–Bloomington
Omitted Responses

• Represent items skipped over by choice, rather than not presented or not reached

• Introduce bias when ignored in 2P and 3P models
  - (Lord, 1974, 1983; Mislevy & Wu, 1988, 1996; Schafer & Graham, 2002)

• Must be filled in with scores
  - What is your best guess?
Four Methods

- Assume response would have been wrong
  - Standard practice
- Assume response would have been consistent with student ability and item difficulty
- Assign average possible score
- Live with the bias
Standard Practice

• Assume response would have been wrong
  • Score as incorrect (constructed response) or fractionally-correct (multiple choice)
  • Assumes student rationally and accurately determined they would get the item wrong
  • Based on very little empirical evidence (Stocking, Eignor, & Cook, 1988; Wang, Wainer, & Thissen, 1995)
Recent Research

- Student ability and item difficulty fail to explain substantial variance in rate of omission
  - Based on a study of the 2009 NAEP Math Assessment (Brown, Dai, & Svetina, 2014)
  - Ability explains 3% of variance in student tendency to omit responses
  - Difficulty explains 5% of variance in item tendency to discourage responses
  - Other predictors also fail to explain substantial variance, with the exception of item format
Recent Research

- Omitted responses appear to be a form of differential item functioning
  - We hypothesize a student skips an item due to unfamiliarity, despite having sufficient ability
  - On the one hand, they likely would have gotten the item wrong
  - On the other hand, that doesn’t appear to reflect on their ability
Alternative Method I

• Assume response would have been consistent with student ability and item difficulty
  • Impute score using available information

• Several imputation procedures exist
  • See Finch (2008) and Schafer & Graham (2002) for reviews
**Alternative Method 2**

- Assign average possible score
- Limits maximum possible error
- Proposed by De Ayala, Plake, & Impara (2001)

<table>
<thead>
<tr>
<th>Possible Scores</th>
<th>Average Possible Score</th>
<th>Maximum Possible Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0, 1, 2</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>0, 1, 2, 3</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>0, 1, 2, 3, 4</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Alternative Method 3

- Live with the bias
  - Treat as not presented
Does the Method Matter?

- Are there substantial differences in NAEP scale scores when different methods are used?
- Some groups omit responses at a higher rate
  - Men omit more frequently than women
  - Blacks and Hispanics omit more frequently than Whites
- Would a different method affect reported group means and achievement gaps?
Data

• 2009 Grade 4 NAEP Mathematics Assessment
  • Responses to the 159 items
  • \( N = 168,850 \) students who responded to at least one item
  • 2% of responses omitted overall
  • 30% of students skipped at least one item
  • 98% of items were skipped at least once
Methodology

• Attempt to reproduce reported NAEP results, starting from scratch

• Generate group means using each of the four methods

• Compare resulting achievement gaps
Methodology

- Use item responses to estimate item parameters
- Define variables for the population-structure model
- Estimate population-structure model parameters
- Estimate group means
- Transform scale scores into reporting metric
NAEP vs. Our Methodology

• Use item responses to estimate item parameters
  • 3PL model for multiple-choice items
  • 2PL model for dichotomous constructed-response items
  • GPC model for polytomous constructed-response items
  • MML estimation
• Using specialized in-house BILOG/PARSCALE software
• Using AM software distributed by NCES
NAEP vs. Our Methodology

- Define variables for the population-structure model
  - PCA on 471 variables from Student, Teacher, and School Questionnaires & 320 variables from the National Indian Education Study
  - Principal components chosen that account for 90% of variance of 18 subgroup variables
  - 3172 original contrasts
  - 1999 original contrasts, 675 principal components
• Estimate population-structure model parameters
  • MML estimation
  • Using in-house CGROUP software
  • Using DESI software
NAEP vs. Our Methodology

- Estimate group means
  - Generate five scale score PVs for each content subscale for each respondent
  - Using CGROUP
  - Using DESI
  - Use PVs to estimate group mean sub-scale scores, accounting for case weights and using jackknife sampling to estimate variances
    - Using CGROUP
    - Using AM
• Transform scale scores into reporting metric
  • Apply NAEP-defined linear transformations to group mean sub-scale scores
  • Calculate composite group mean scale scores using NAEP-defined sub-scale weights
## Reproducing NAEP Results

### Item Parameter Estimation

<table>
<thead>
<tr>
<th>Sub-Scale</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Properties &amp; Operations</td>
<td>Bias</td>
<td>–0.10</td>
<td>–0.11</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.15</td>
<td>0.21</td>
</tr>
<tr>
<td>Measurement</td>
<td>Bias</td>
<td>–0.08</td>
<td>–0.01</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.14</td>
<td>0.26</td>
</tr>
<tr>
<td>Geometry</td>
<td>Bias</td>
<td>0.01</td>
<td>–0.05</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.07</td>
<td>0.20</td>
</tr>
<tr>
<td>Data Analysis, Statistics &amp; Probability</td>
<td>Bias</td>
<td>0.10</td>
<td>–0.02</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.41</td>
<td>0.11</td>
</tr>
<tr>
<td>Algebra</td>
<td>Bias</td>
<td>–0.00</td>
<td>–0.03</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.04</td>
<td>0.07</td>
</tr>
</tbody>
</table>
M154901  (F=2.584277, p>F=0.088865)
Reproducing NAEP Results

### Item Parameter Estimation

<table>
<thead>
<tr>
<th>Sub-Scale</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Properties &amp; Operations</td>
<td>–0.10</td>
<td>–0.11</td>
<td>–0.08</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.15</td>
<td>0.21</td>
<td>0.09</td>
</tr>
<tr>
<td>Measurement</td>
<td>–0.08</td>
<td>–0.01</td>
<td>–0.05</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.14</td>
<td>0.26</td>
<td>0.07</td>
</tr>
<tr>
<td>Geometry</td>
<td>0.01</td>
<td>–0.05</td>
<td>–0.04</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.07</td>
<td>0.20</td>
<td>0.09</td>
</tr>
<tr>
<td>Data Analysis, Statistics &amp; Probability</td>
<td>0.00</td>
<td>–0.03</td>
<td>–0.03</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.08</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>Algebra</td>
<td>–0.00</td>
<td>–0.03</td>
<td>–0.03</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.04</td>
<td>0.07</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Excluding Item M154901
Reproducing NAEP Results

Item Parameter Estimation

• Differences not due to estimation stability
  • Using other initial values, including the NAEP item parameters, led to very similar item estimates

• Most likely due to differences in the estimation software
  • BILOG/PARSCALE vs. AM

• Close enough
## Reproducing NAEP Results

### Group Mean Estimation

<table>
<thead>
<tr>
<th>Group</th>
<th>Weighted N</th>
<th>NAEP</th>
<th>Our</th>
<th>SE</th>
<th>Δ</th>
<th>NAEP</th>
<th>Our</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1,901,423</td>
<td>240.6</td>
<td>241.1</td>
<td>0.3</td>
<td>0.5</td>
<td>29.5</td>
<td>30.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Female</td>
<td>1,839,818</td>
<td>238.7</td>
<td>239.0</td>
<td>0.3</td>
<td>0.3</td>
<td>27.9</td>
<td>29.3</td>
<td>1.4</td>
</tr>
<tr>
<td>White</td>
<td>2,080,531</td>
<td>248.1</td>
<td>248.9</td>
<td>0.2</td>
<td>0.8</td>
<td>25.5</td>
<td>26.8</td>
<td>1.2</td>
</tr>
<tr>
<td>Black</td>
<td>592,861</td>
<td>222.3</td>
<td>221.8</td>
<td>0.3</td>
<td>−0.6</td>
<td>26.8</td>
<td>28.1</td>
<td>1.2</td>
</tr>
<tr>
<td>Hispanic</td>
<td>780,823</td>
<td>227.5</td>
<td>227.1</td>
<td>0.4</td>
<td>−0.3</td>
<td>27.2</td>
<td>28.6</td>
<td>1.4</td>
</tr>
<tr>
<td>Asian Am.</td>
<td>184,817</td>
<td>254.8</td>
<td>256.1</td>
<td>1.0</td>
<td>1.3</td>
<td>29.1</td>
<td>30.4</td>
<td>1.3</td>
</tr>
<tr>
<td>Am. Indian</td>
<td>42,742</td>
<td>224.9</td>
<td>225.7</td>
<td>0.9</td>
<td>0.8</td>
<td>29.3</td>
<td>28.6</td>
<td>−0.7</td>
</tr>
</tbody>
</table>
Reproducing NAEP Results

Group Mean Estimation

- Differences partly due to the differences in the population-structure model
- Probably mostly due to differences in drawing and analyzing PVs
  - CGROUP vs. DESI
- Close enough
## Reproducing NAEP Results

### Overall Results

<table>
<thead>
<tr>
<th>Group</th>
<th>Achievement Gap</th>
<th>Effect Size d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Target</td>
<td>Comparison</td>
</tr>
<tr>
<td>Female</td>
<td>Male</td>
<td>–1.9</td>
</tr>
<tr>
<td>Black</td>
<td>Male</td>
<td>–25.7</td>
</tr>
<tr>
<td>Hispanic</td>
<td>Male</td>
<td>–20.6</td>
</tr>
<tr>
<td>Asian Am.</td>
<td>White</td>
<td>6.7</td>
</tr>
<tr>
<td>Am. Indian</td>
<td>Male</td>
<td>–23.2</td>
</tr>
<tr>
<td>ELL</td>
<td>Non-ELL</td>
<td>–24.4</td>
</tr>
<tr>
<td>Lunch eligible</td>
<td>Not eligible</td>
<td>–22.6</td>
</tr>
<tr>
<td>IEP</td>
<td>Non-IEP</td>
<td>–21.2</td>
</tr>
<tr>
<td>Private</td>
<td>Public</td>
<td>6.9</td>
</tr>
</tbody>
</table>
Methodology

• Attempt to reproduce reported NAEP results, starting from scratch

• Generate group means using each of the four methods

• Compare resulting achievement gaps
Details on Four Methods

- **Standard practice ("Incorrect")**
  - Implemented directly in AM and DESI

- **Treat as not presented ("Ignore")**
  - Implemented directly in AM and DESI

- **Assign average possible score ("Average")**
  - AM and DESI do not accept non-integer scores
  - When average is non-integer (e.g., 1.5), adjacent integer scores (e.g., 1, 2) were assigned randomly with equal probability
Details on Methods

• Imputation ("Impute")
  • Following Finch (2008)
  • Multiple imputation using SAS PROC MI, generating 5 datasets
  • Imputed scores rounded to nearest valid integer score
  • Item parameters estimated from each dataset, resulting parameters averaged together
  • Plausible values generated using each dataset, resulting scale scores averaged together
## Comparison of Methods

<table>
<thead>
<tr>
<th>Group</th>
<th>Target Comparison</th>
<th>Incorrect</th>
<th>Ignore</th>
<th>Average</th>
<th>Impute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Male</td>
<td>–0.07</td>
<td>–0.09</td>
<td>–0.09</td>
<td>–0.09</td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>–0.99</td>
<td>–0.98</td>
<td>–0.98</td>
<td>–0.97</td>
</tr>
<tr>
<td>Hispanic</td>
<td>White</td>
<td>–0.79</td>
<td>–0.79</td>
<td>–0.79</td>
<td>–0.79</td>
</tr>
<tr>
<td>Asian Am.</td>
<td></td>
<td>0.28</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>Am. Indian</td>
<td></td>
<td>–0.85</td>
<td>–0.85</td>
<td>–0.84</td>
<td>–0.84</td>
</tr>
<tr>
<td>ELL</td>
<td>Non-ELL</td>
<td>–0.86</td>
<td>–0.87</td>
<td>–0.86</td>
<td>–0.86</td>
</tr>
<tr>
<td>Lunch eligible</td>
<td>Not eligible</td>
<td>–0.85</td>
<td>–0.85</td>
<td>–0.85</td>
<td>–0.85</td>
</tr>
<tr>
<td>IEP</td>
<td>Non-IEP</td>
<td>–0.75</td>
<td>–0.75</td>
<td>–0.75</td>
<td>–0.74</td>
</tr>
<tr>
<td>Private</td>
<td>Public</td>
<td>0.23</td>
<td>0.24</td>
<td>0.23</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Conclusions

• For group comparisons with very large N, different methods are indistinguishable
  • Largest effect seen on gender gap, $\Delta d = -0.02$

• Little incentive to change methodology for national and international assessments that focus on group comparisons
  • e.g., NAEP, TIMSS, PIRLS, IALS
Gains for Ommitters
When Omitted Responses are Ignored

Mean Change
3.3 ($d = 0.11$)
Change of $d > 0.2$
16%

Maximum Change
61.7 ($d = 2.08$)
Losses for Non-Omitters
When Omitted Responses are Ignored

Mean Change
\(-0.7 (d = -0.02)\)
Conclusions

- For individuals, alternative methods lead to modest to large gains for omitters
- Non-omitters suffer negligible losses
- For smaller group comparisons, alternative methods may lead to different results
  - Comparing classrooms or teachers, in particular
Conclusions

• Shouldn’t use standard practice just because it’s standard practice
  • Assumption that omitted responses reflect poorly on omitters appears to be shaky

• Different alternative methods give similar results

• Ignoring omitted responses is the easiest to implement
  • As easy as standard practice
  • If rate of omission is not strongly related to ability, bias may be minimal