Score Resolution in Essay Grading:

A View from a Signal Detection Model of Rater Behavior

Lawrence T. DeCarlo
Teachers College, Columbia University

and

YoungKoung Kim
The College Board
What’s the Purpose of a Scoring Rubric?

• One view:
  • The scoring rubric defines latent classes of essays
  • Tasks of raters are to discriminate between the latent classes

• How do raters score essays?
  • Signal Detection Theory (SDT):
    • SDT provides a psychological theory of the behavior of raters
    • Has been widely and successfully used in psychology and medicine
    • Raters have a perception of the quality of each essay
    • Perception is used in conjunction with response criteria \((cj)\) to arrive at a response
An example with a 1 to 4 response and 4 latent classes:

- $d$ indicates a rater’s ability to discriminate between the latent classes – primary rater parameter of interest
- $c$ reflects the arbitrary response tendencies of raters, such as being lenient or being strict
An example with a 1 to 4 response and 4 latent classes (con’t):

• $d$ has a large effect on classification accuracy, $c$ has only a small effect
  • Suggests that attempts to improve agreement may not be cost effective (if it is due to differences in $c$)
• Agreement reflects both $c$ and $d$
  • Can have excellent discrimination but poor agreement, because of differences in $c$

Adjudication

- Typically two raters per essay in large scale assessments
- If the scores of raters differ by more than one point, a third score is obtained
- SDT provides a theoretical framework for addressing issues pertaining to adjudication
  - Simulations can be used to examine the effects of adjudication on classification accuracy
  - Analysis of real world data can be used to evaluate adjudicated scores
Simulation

• Data generated according to latent class SDT model
  • Sample size of 20,000
  • 100 replications
• Parameters similar to those found for a number of datasets (d’s of 3, 3.5)
• Data for three raters were generated
  • For the third rater, missing values were substituted in cases where the difference between the first two raters was less than one
• Simulation Results
  • The percentage of cases that needed to be adjudicated were the same as that of those found for real world data
    • The percent of adjudicated cases is consistent with the SDT model
    • Thus, the differences do not necessarily indicate a problem with the rater and/or essay
Simulation Results

• Simulation Results: Classification
  • Overall classification accuracy (PC) was 71% (e.g., using the average of the two scores)
  • For adjudicated cases (8% of total), PC was 69%
    • Only 2% lower than overall PC
  • For adjudicated cases, using a third score raised the PC to 74%
    • In terms of fairness, essays with three raters have a classification accuracy that was higher than that of the overall
  • Using the third score alone or an average with closest score gave the lowest PC, about 64%

• Simulation Results: Estimation
  • Non-adjudicated cases are missing at random for the third rater
  • Estimation with 92% missing data for the third rater was excellent (n = 20,000)
**Results for SAT**

- Random subsample of 20,000 from two administrations
- 3-4% of cases required adjudication

<table>
<thead>
<tr>
<th></th>
<th>Administration 1</th>
<th>Administration 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimate</strong></td>
<td><strong>SE</strong></td>
<td><strong>Estimate</strong></td>
</tr>
<tr>
<td>$d_1$</td>
<td>3.46</td>
<td>0.08</td>
</tr>
<tr>
<td>$d_2$</td>
<td>3.60</td>
<td>0.09</td>
</tr>
<tr>
<td>$d_3$</td>
<td>3.66</td>
<td>0.46</td>
</tr>
</tbody>
</table>

- Discrimination for the adjudicated scores, $d_3$, was equal to $d_1$ and $d_2$
- SE’s are larger because 96% missing data
Conclusions

• No evidence that adjudicated cases differ from other cases
• Classification accuracy for adjudicated cases is only slightly smaller
  • Raises questions about the cost effectiveness of adjudication
• Present approach offers a way to quantitatively evaluate effects of adjudication
  • Using a third rater led to about a 5% increase for the current example
• Can also evaluate assumptions about the raters or gold standards

Thank you!