Analysis of Norm-Referencing Modifiers as a Component of Rating Rubrics

Alex Volkov, Jake Stone, Taylor Asbury, Michael Hsu
Paragon Testing Enterprises
University of British Columbia
Agenda

✓ Rationale
✓ Research questions
✓ Methodology
✓ Analysis
✓ Conclusion
Rating Scale Design

✓ Empirically informed rating rubric
✓ How do modifiers function as a component of rating rubrics?
✓ Lack of research in modifiers
Research Questions

✓ Do modifiers correspond to a specific ability range?

✓ Will the same modifier attached to different descriptors nonetheless be targeting a similar ability range?
Descriptors and Modifiers

✓ Creates a cohesive text
✓ Grammar is correct
✓ Punctuation is correct
✓ Vocabulary is appropriate
✓ Writing is intelligible

✓ Partially
✓ Sufficiently
✓ Mostly

Creates a **partially** cohesive text
Creates a **sufficiently** cohesive text
Creates a **mostly** cohesive text
Methodology

✓ 30 Samples, 10 experienced raters, 5 descriptor types, 3 modifiers. Fully crossed design

✓ Online rating using Fluid Surveys
Test Taker's Response

Dear Sir(Mr.x),

Kindly, my son Albert is one of your student, in the 6th grade class at North elementary school. I am writing you today to express my concerns, regarding Albert feelings during the last semester. As You know, the academic program for this year class is charged with technical scientific subject and the children have to prepare and present a lot of projects. Although, Albert was always interested in the material you presented, he started feeling uncomfortable about the project handling. He expressed a lot of anxious regarding the subject presentation, and his colleagues behaviour during the presentation time, especially the interaction, questions and answer parts. Despite putting a lot of effort in the research and preparation, Albert is always afraid to present his job, in front of audience. I was trying to help him developing more confidence, by making him do more practices. If would appreciate if we can meet together, to discuss the best way to follow in order to help Albert overcome his fears.

Your Input regarding the above mentioned issue, will be highly appreciated.

Regards

Check the descriptor below if the response meets the ability described by this descriptor

- [ ] Vocabulary is mostly appropriate
Analysis

✓ Descriptive statistics

✓ ANOVA

✓ MFRM
Analysis: Descriptive Statistics

✓ Creates a **partially** cohesive text

✓ Creates a **sufficiently** cohesive text

✓ Creates a **mostly** cohesive text
Analysis: Descriptive Statistics

✓ Grammar is **partially** correct

✓ Grammar is **sufficiently** correct

✓ Grammar is **mostly** correct
Analysis: Box Plots
(across all descriptors)
Analysis: ANOVA

<table>
<thead>
<tr>
<th>Descriptor type</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptor</td>
<td>28</td>
<td>4</td>
<td>7.0</td>
<td>2.718</td>
<td>0.028419</td>
</tr>
<tr>
<td>Modifier</td>
<td>1945</td>
<td>2</td>
<td>972.5</td>
<td>376.471</td>
<td>0.000000</td>
</tr>
<tr>
<td>Descriptor * Modifier</td>
<td>82</td>
<td>8</td>
<td>10.2</td>
<td>3.963</td>
<td>0.000117</td>
</tr>
<tr>
<td>Residuals</td>
<td>3895</td>
<td>1508</td>
<td>2.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

✓ Descriptor type
✓ Modifier
✓ Statistical significance
Analysis: ANOVA
Analysis: MFRM

- Grammar is mostly correct
- Punctuation is mostly correct
- Creates a mostly cohesive text
- Grammar is sufficiently correct
- Punctuation is sufficiently correct
- Creates a sufficiently cohesive text
- Writing is mostly intelligible
- Vocabulary is mostly appropriate
- Writing is sufficiently intelligible
- Vocabulary is sufficiently appropriate
- Grammar is partially correct
- Creates a partially cohesive text
- Vocabulary is partially appropriate
- Writing is partially intelligible
# Analysis. MFRM

## Table 7.1.1 Descriptors Measurement Report (arranged by MN).

<table>
<thead>
<tr>
<th>Total Score</th>
<th>Total Count</th>
<th>Obsvd Average</th>
<th>Error(M) Average</th>
<th>Error Measure</th>
<th>Infit Mnsq Zstd</th>
<th>Oufit Mnsq Zstd</th>
<th>Estim. of Discrim</th>
<th>Correlation PtMea PtExp</th>
<th>Nu Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>719</td>
<td>85</td>
<td>8.46</td>
<td>8.49</td>
<td>.54</td>
<td>.08</td>
<td>.98</td>
<td>1.02</td>
<td>.1</td>
<td>.95</td>
</tr>
<tr>
<td>1110</td>
<td>141</td>
<td>7.87</td>
<td>7.86</td>
<td>.27</td>
<td>.05</td>
<td>1.11</td>
<td>1.0</td>
<td>1.12</td>
<td>1.1</td>
</tr>
<tr>
<td>842</td>
<td>107</td>
<td>7.87</td>
<td>7.82</td>
<td>.26</td>
<td>.06</td>
<td>1.13</td>
<td>1.1</td>
<td>1.14</td>
<td>1.1</td>
</tr>
<tr>
<td>1045</td>
<td>135</td>
<td>7.74</td>
<td>7.72</td>
<td>.22</td>
<td>.05</td>
<td>.99</td>
<td>1.00</td>
<td>1.0</td>
<td>.97</td>
</tr>
<tr>
<td>1063</td>
<td>142</td>
<td>7.49</td>
<td>7.48</td>
<td>.14</td>
<td>.05</td>
<td>1.01</td>
<td>1.0</td>
<td>1.02</td>
<td>1.1</td>
</tr>
<tr>
<td>844</td>
<td>113</td>
<td>7.47</td>
<td>7.44</td>
<td>.13</td>
<td>.05</td>
<td>.69</td>
<td>-3.4</td>
<td>.69</td>
<td>3.4</td>
</tr>
<tr>
<td>786</td>
<td>118</td>
<td>6.66</td>
<td>6.63</td>
<td>-.13</td>
<td>.06</td>
<td>1.22</td>
<td>2.2</td>
<td>1.21</td>
<td>2.1</td>
</tr>
<tr>
<td>726</td>
<td>112</td>
<td>6.48</td>
<td>6.41</td>
<td>-.20</td>
<td>.05</td>
<td>.93</td>
<td>-.7</td>
<td>.92</td>
<td>-.7</td>
</tr>
<tr>
<td>701</td>
<td>110</td>
<td>6.37</td>
<td>6.27</td>
<td>-.25</td>
<td>.06</td>
<td>1.05</td>
<td>-.5</td>
<td>1.06</td>
<td>-.5</td>
</tr>
<tr>
<td>645</td>
<td>104</td>
<td>6.20</td>
<td>6.11</td>
<td>-.30</td>
<td>.06</td>
<td>.93</td>
<td>-.6</td>
<td>.91</td>
<td>-.7</td>
</tr>
<tr>
<td>407</td>
<td>71</td>
<td>5.73</td>
<td>5.75</td>
<td>-.44</td>
<td>.08</td>
<td>1.14</td>
<td>1.9</td>
<td>1.16</td>
<td>1.0</td>
</tr>
<tr>
<td>321</td>
<td>65</td>
<td>4.94</td>
<td>4.95</td>
<td>-.90</td>
<td>.11</td>
<td>1.13</td>
<td>1.5</td>
<td>1.11</td>
<td>1.5</td>
</tr>
<tr>
<td>501</td>
<td>102</td>
<td>4.91</td>
<td>4.87</td>
<td>-.94</td>
<td>.09</td>
<td>.81</td>
<td>-.9</td>
<td>.77</td>
<td>1.1</td>
</tr>
<tr>
<td>291</td>
<td>63</td>
<td>4.62</td>
<td>4.59</td>
<td>-.12</td>
<td>.14</td>
<td>.79</td>
<td>-.6</td>
<td>.75</td>
<td>-.8</td>
</tr>
<tr>
<td>253</td>
<td>55</td>
<td>4.60</td>
<td>4.57</td>
<td>-.12</td>
<td>.16</td>
<td>.77</td>
<td>-.6</td>
<td>.79</td>
<td>-.5</td>
</tr>
</tbody>
</table>

Model: Populn: RMSE .08 Adj (True) S.D. .55 Separation 6.61 Strata 9.14 Reliability .98
Model: Sample: RMSE .08 Adj (True) S.D. .57 Separation 6.84 Strata 9.46 Reliability .98
Model, Fixed (all same) chi-square: 532.2 d.f.: 14 significance (probability): .00
Model, Random (normal) chi-square: 13.5 d.f.: 13 significance (probability): .41
Feedback from Raters

✓ Even though the raters were not told the purpose of the study, they noticed the systematic use of the modifiers
✓ Many raters informed us that limiting the number of modifiers to 3 was very helpful
✓ Each rater began to devise heuristics for judging each modifier
Future work

✓ Can training or seminars help raters develop a shared consensus as to how different performance levels correspond with different modifiers?

✓ Can we discern a limited and highly descriptive pool of modifiers for systematic application?
Conclusion

✓ Modifiers can be discerning and can have a major effect on raters’ perceptions

✓ Modifiers should be carefully selected. Some are better targeted at a specific ability range.

✓ Using a limited number of modifiers systematically may help with inter-rater reliability
Thank you!
Analysis of Norm-Referencing Modifiers as a Component of Rating Rubrics

Questions?