Diagnostic Item Analysis using R:

John Denbleyker
Minnesota Department of Education

Shuqin Tao
Data Recognition Corporation

CCSSO 2012 National Conference on Student Assessment
Minneapolis, MN
June, 2012
The example Diagnostic Item Graphic is intended to empirically display strengths and weaknesses based on item response data at an aggregate level (e.g., schools, student group and intervention programs) utilizing a robust effect-size measure estimated via Bayesian methods.
PRESENTATION ORDER

♦ Literature on graphic comprehension of data/visualization and score reporting
♦ Examples of aggregate item-level score reports
♦ Overview of Diagnostic Item Analysis graphic
♦ Comparison of Methods
♦ Summary
SCORE REPORTING = DATA VISUALIZATION

Graphical comprehension of quantitative data

- Wainer (1997, 2005)
- Cleveland (1985)

The above literature provides the benefit of years of research on the best way to

1) represent data, from color schemes that convey meaning to creating elegant displays that keeps users focused on what is essential.

2) To create visualizations of test score data in order to assist educational stakeholders understand and communicate information in the most effective way possible.
SCORE REPORTING = DATA VISUALIZATION

Gallery of Data Visualization
http://www.datavis.ca/gallery/

Many Eyes (IBM Research and the IBM Cognos software group)

googleVis R package

...good graphical displays of data communicate ideas with clarity, precision, and efficiency.

...bad graphical displays distort or obscure the data
The paper discussed three possible concerns with many score reporting practices.
1st concern of information presented in graphs potential over-reporting of underspecified or unreliable constructs (Twing, 2008)

♦ Graphical depictions can be highly salient
⇒ making some information more visually salient than other information may lead to additional interpretation errors.
2nd concern of information presented in graphs

score reports can disregard the different goals and abilities of the intended audience.

— individual differences in statistical and quantitative understanding of graphs and data
— individual differences in prior knowledge and dispositions (willingness to include collateral information)

These two components can have an impact on the interpretation of any score report.
3rd concern of information presented in graphs

Score report users may not critically evaluate information offered, but instead focus on one or two salient bits of information
Some additional conclusions from the literature as specified by the researchers

The comprehension of graphs is not just affected by numeracy skills but is also substantially knowledge-driven – information that is more difficult to process is actually better understood and remembered – static displays require active processing, such as mental animation, display animations are more likely to be viewed passively (Hegarty, 2004; Hegarty, Kriz, & Cate, 2003) – Data presented in less-processed formats may lead to more initial difficulty in comprehension, but also fewer misinterpretations

When data depicted in graphics are familiar to viewers, they were better able to draw appropriate inferences When data were unfamiliar, people primarily focused on salient visual information (Shah & Freedman, 2009).
RYAN (2003; 2006)

An analysis of item mapping and test reporting strategies. Greensboro, NC: SERVE


“The review of research on score reports shows that many practitioners find the use of graphical displays helpful in interpreting test results. The graphical displays in most score reports, however, are fairly conventional and are used to convey such basic information as number or percent of item correct, scale scores, or scale means”
Student test score reports and interpretive guides: Review of current practices and suggestions for future research. Applied Measurement in Education, 17(2), 145-220
Efforts to Produce Relevant Score Reports to School, District, and State Officials on National Tests

They noted the following:

♦ As assessment developers, it is our responsibility to ensure that the information displayed on score reports are understood, meaningful, and useful to their intended audiences.

• Standards for Educational and Psychological Testing
• Code of Fair Testing Practices in Education
• Code of Professional Responsibilities in Educational Measurement
Designing and Evaluating Score Reports for Particular Audiences

“Although principles for designing high-quality score reports have been proposed and professional standards indicate that test takers need to be informed about assessment results as well as the purpose of the assessment and its recommended uses, many of the score reports available do not effectively convey this score information for particular audiences.”

This research set to design and evaluate score reports to clearly communicate useful assessment information to various educational stakeholders (e.g., teachers, administrators, and students)
ROBERTS AND GIERL (2010)


- reviews current score reporting practices
- proposed a framework for developing score reports
- highlight the importance of evaluating the score reports with the intended audience.
"...provides an overview of score reports from a development perspective, focusing on current practices and emerging efforts in content of reports as well as the process by which reports are designed, evaluated, and ultimately used to communicate with the public."
From “Here’s the Story” to “You’re in Charge”: Development and maintaining large-scale online test and score reporting resources. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), Handbook of large-scale assessment. London: Taylor & Francis

Framework for Designing and Evaluating Score Reports

They present an extensive list of score reporting resources that includes papers, guidelines, and sample score reports.
TWO EXAMPLES

Useful Information

Limited Visualization

Statistical Inferential Issues
### Proportion Correct on Items by School

<table>
<thead>
<tr>
<th>School</th>
<th>Raw Score</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rachel Alexandra</td>
<td>39.72</td>
<td>0.72</td>
<td>0.79</td>
<td>0.98</td>
<td>0.46</td>
<td>0.80</td>
<td>0.83</td>
<td>0.97</td>
<td>0.86</td>
<td>0.74</td>
<td>0.56</td>
</tr>
<tr>
<td>n = 218</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zenyetta</td>
<td>34.63</td>
<td>0.55</td>
<td>0.74</td>
<td>0.90</td>
<td>0.81</td>
<td>0.43</td>
<td>0.70</td>
<td>0.90</td>
<td>0.53</td>
<td>0.51</td>
<td>0.30</td>
</tr>
<tr>
<td>n = 317</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Curlin</td>
<td>33.47</td>
<td>0.58</td>
<td>0.61</td>
<td>0.89</td>
<td>0.67</td>
<td>0.42</td>
<td>0.82</td>
<td>0.84</td>
<td>0.70</td>
<td>0.59</td>
<td>0.54</td>
</tr>
<tr>
<td>n = 265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sea The Stars</td>
<td>35.70</td>
<td>0.48</td>
<td>0.59</td>
<td>0.92</td>
<td>0.66</td>
<td>0.56</td>
<td>0.72</td>
<td>0.95</td>
<td>0.87</td>
<td>0.44</td>
<td>0.31</td>
</tr>
<tr>
<td>n = 170</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality Road</td>
<td>35.51</td>
<td>0.63</td>
<td>0.55</td>
<td>0.90</td>
<td>0.66</td>
<td>0.51</td>
<td>0.66</td>
<td>0.93</td>
<td>0.67</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>n = 253</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>35.57</td>
<td>0.59</td>
<td>0.66</td>
<td>0.91</td>
<td>0.67</td>
<td>0.53</td>
<td>0.74</td>
<td>0.91</td>
<td>0.70</td>
<td>0.55</td>
<td>0.44</td>
</tr>
<tr>
<td><strong>n = 1280</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CLASSIFICATION OF CURRICULAR STRENGTHS/WEAKNESSES

Example:
IRT residual-based
Mean of differences between predicted probability (based on IRT model and total ability level) versus actual (0/1) response
Avg. school residual +/- s.e.
Strength/Weakness/Equal
# IRT Residual Approach

## Geometry & Measurement

Describe and classify three-dimensional figures including cubes, prisms and pyramids by the number of edges, faces or vertices as well as the types of faces.

Recognize and draw a net for a three-dimensional figure.

Develop and use formulas to determine the area of triangles, parallelograms and figures that can be decomposed into triangles.

Use various tools and strategies to measure the volume and surface area of objects that are shaped like rectangular prisms.

Understand that the volume of a three-dimensional figure can be found by counting the total number of same-sized cubic units that fill a shape without gaps or overlaps. Use cubic units to label volume measurements.

Develop and use the formulas \( V = lwh \) and \( V = Bh \) to determine the volume of rectangular prisms. Justify why base area \( B \) and height \( h \) are multiplied to find the volume of a rectangular prism by breaking the prism into layers of unit cubes.
LIMITATIONS

p-value by school chart lacks:
1) a comparative graphic
2) a base metric
3) lacks a direct inferential effect-size metric
   (e.g., a statistical significance test for p-value differences; however, can bootstrap sample)

IRT residual classification lacks:
1) a normative comparison
2) a comparative graphic
3) a standard reference of comparison
   (e.g., each school is their own reference as compared to a state-wide mean)
A relative effect size that is graphically portrayed via a forest-type plot—compare to a reference(s) group

- incorporates:
  - faceting (same graph over subsets)
  - credible intervals
  - types of items (benchmark, standard, etc)
  - absolute comparison
  - relative comparison
  - various sorting of data to enhance comparisons
PRACTICAL USES

The diagnostic item analysis graph can assist teachers, curriculum leaders, and directors of research/assessment departments in the analysis of test results at the item or standard level.

Information provided is in a relative manner (e.g. comparing one group to another) and provides a common metric that makes comparisons of strengths and weaknesses of items/standard more readily apparent.
COMMON LANGUAGE EFFECT SIZE STATISTIC (CLES) MCGRAW AND WONG (1992)

\[ CLES = P(X_1 > X_2) = \Phi \left( \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \right) \]

probability that a randomly selected student from one group will get the item correct compared to a randomly sampled student from the other group
PROCESS OF GRAPHIC PRODUCTION

Software: R
♦ BRugs/JAGS/learnBayes/

Bayesian estimation of effect-size
♦ ggplot2

Graphical production
DIAGNOSTIC ITEM ANALYSIS FOR STATE SCIENCE ASSESSMENT

Data
34-item G5 and G8 Science Test (MC items only)

MCA-II Science Assessment
Data from a school in a large MN district
2 items on same standard consistent

2 items on same standard NOT consistent
DIAGNOSTIC ITEM ANALYSIS GRAPHIC OFFERS

→ Compare items in a robust and scale-free manner
→ Compare credible intervals for meaningful differences between p-values
→ Compare different tests on the same metric
→ Visual graphic that facilitates understanding the context of the data
→ Intuitive measure of effect-size built into the graphical display
USES OF DIAGNOSTIC ITEM ANALYSIS GRAPHIC

→ Compare items to state/district/school/specific student group in an **absolute** fashion
→ Compare items to one another in a **relative** fashion
→ Program evaluation (compare intervention vs. non-intervention group)
→ Methodology can be used with distractors as well to garner diagnostic information on student misunderstanding and areas of curricular/teaching weakness
→ Use in conjunction with absolute measures (p-value) for analysis utilizing both normative and absolute measure of effectiveness and achievement


