Establishing baseline data for incidents of misconduct in the next generation assessment environment

Deborah J. Harris, Chi-Yu Huang, and Rya Dunnington
ACT, Inc.
What (really) is benchmarking?

Benchmarking, a systematic comparison of the processes and practices of two or more companies or two or more units of a company, gauges the performance of an organization or unit relative to a peer.

(http://hbswk.hbs.edu/archive/3746.html)
Why is benchmarking “better” than statistical probability?

Consider looking at the responses two examinees have in common on a multiple choice test.

Statistically, one could argue that that the examinees have \( \frac{1}{4} \times \frac{1}{4} \) chance of both selecting option “a” on item 1 randomly.

BUT Are they selecting randomly?

Are they selecting independently?

Is option “a” more or less attractive?
Benchmark data provides a richer context in which to interpret results. It also serves as a way to “account for” variables and conditions we can’t really account for, or can’t account for easily.
For example:

We “know” most examinees do not answer questions randomly. However, when we are looking at how similar two sets of examinee responses are, we don’t know what the tendency is for examinees who answer “c” to Item 7 to answer “d” to Item 8. Benchmark data can take this tendency into account, though indirectly.
Consider two pairs of examinees, one pair has 40 item responses in common; the other pair has 24 item responses in common.

We then look at additional information, such as how many of those common responses are incorrect.

Then we look at those pairs in relation to benchmark data.
Response Similarity

The graph shows the relationship between the number of same incorrect responses and the number of same responses. The data points are distributed along a scatter plot, indicating a trend where as the number of same incorrect responses increases, the number of same responses also tends to increase. Two red arrows highlight specific points on the graph, possibly indicating significant or notable data points.
There are different ways to compute baseline data.

One might be considered, in a sense, a null case.

For example, for developing a copier set of baseline data, examinees who test in different states may be paired to determine how many like responses they have.
Other benchmark data might be considered, in a sense, less pure but more relevant.

For example, comparing all examinee pairs within a particular district, whether they tested in the same room or not, to have the “same curriculum” as a controlled factor.
We might initially run benchmark statistics by individual test form, or test date. But later, collapse across these and have a single set of benchmark data.

It’s a balance between the context (this form, this set of co-examinees), and standardization (this is the single constant rule to flag any examinee, across forms and across examinee cohorts)
There is always someone most extreme.

There are always “outliers”.

The question is, does it matter in our context?
Response Similarity/String
Having baseline data to compare values to both increases the likelihood of the values not being over or under interpreted, and provides a context in which to present information if the decision is made to pursue an incident.

But how we display the information may also matter.....
Erasure Patterns

Percentage

Items displayed by item difficulty (from hard to easy)

Others

W to R

No erasures

ACT
improveyourself.org
Erasure Patterns

![Graph showing erasure patterns with different items displayed by item sequence.](image_url)
Erasure Patterns: Wrong to Right

Percentage

Item Sequence

TC003
Benchmark
Erasure Patterns: Wrong to Right

<table>
<thead>
<tr>
<th>Item Sequence</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>40</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
</tr>
</tbody>
</table>

- TC001
- Benchmark
As stated earlier, benchmark data can be “refined”, or made more relevant....
Latency Data

![Graph showing latency data with two lines representing First timer and Repeater, with item sequence from 1 to 31 and response latency average in seconds ranging from 0 to 250.]
Latency Data

![Graph showing latency data for different educational levels. The x-axis represents item sequence, and the y-axis represents response latency (seconds). The graph compares Total group, High School, Community College, Trade/professional School, 4-year College, and College Graduate.](image-url)
Option analysis (what percent of students choose each available option on a multiple choice or selection-type item) is often done to identify test site anomalies.
## Item Performance by Test Center

<table>
<thead>
<tr>
<th>TC</th>
<th>option1_item1</th>
<th>option2_item1</th>
<th>option3_item1</th>
<th>option4_item1</th>
<th>option1_item2</th>
<th>option2_item2</th>
<th>option3_item2</th>
<th>option4_item2</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>0</td>
<td>55.56</td>
<td>44.44</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>002</td>
<td>0</td>
<td>16.67</td>
<td>83.33</td>
<td>0</td>
<td>33.33</td>
<td>0</td>
<td>0</td>
<td>66.67</td>
</tr>
<tr>
<td>003</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>004</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16.67</td>
</tr>
<tr>
<td>005</td>
<td>0</td>
<td>14.29</td>
<td>85.71</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>006</td>
<td>0</td>
<td>27.27</td>
<td>72.73</td>
<td>0</td>
<td>9.09</td>
<td>0</td>
<td>0</td>
<td>90.91</td>
</tr>
<tr>
<td>007</td>
<td>3.77</td>
<td>11.32</td>
<td>84.91</td>
<td>0</td>
<td>1.89</td>
<td>0</td>
<td>5.66</td>
<td>92.45</td>
</tr>
<tr>
<td>008</td>
<td>0</td>
<td>32</td>
<td>64</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>12</td>
<td>84</td>
</tr>
</tbody>
</table>
Option Analysis: Correct Option

![Graph showing item sequence and P-values for different groups.](image-url)
Option Analysis: Correct Option

![Graph showing item sequence and p-values for history group, current group, and TC100. The graph plots item p-values against item sequence from easy to hard. The history group and current group show trends in p-values as items become more difficult.]
1 circle/test center representing the percent of examinees at that test center who selected this response.

Quartile distribution indicators

Orange = correct response

Blue = incorrect response

representation of P-value (expressed as 0-100, e.g., 75.20 = p value .752)
Number of Same Responses v. Longest Consecutive String of Same Responses

Baseline Pairs
Test Center 1 Pairs
Longest Consecutive String of Same Responses v. Number of Incorrect Responses in String
Longest Consecutive String of Same Responses v. Number of Same Responses
Thank you