The Usual Suspects: Using Excel to Identify Problematic Items and Response Patterns from Multiple Choice Tests with Small Samples

Jeremy Penn, Ph.D.
Director of Assessment
North Dakota State University
AALHE, 2017
In Memory:
Del Harnisch
1949-2016

Advisor, teacher, friend
What is the most interesting exam cheating incident you’ve experienced?
When you suspect there is an issue with an exam you’ve given in class (say, cheating)…who do you investigate?

I’ve lined up some suspects for you to consider:
Nick Nolte, 2002
Justin Bieber,
2014
Shia LeBeouf, 2014
Shia LeBeouf, 2015
HE WILL NOT DIVIDE US

Shia LeBeouf, 2017
Or was it one of these guys?
Unfortunately,

• No item statistic can catch all cheaters as cheating takes many forms!
• If you are specifically interested in detecting cheating behaviors you might consider research like:
Ways an examinee’s test score can go wrong

- Cheating (copying from someone else answers the examinee would not have been able to answer correctly on his / her own)
- Careless responding (answers certain items wrong that he / she should have gotten correct)
- Lucky guessing (get some correct answers by chance)
- Creative responding (miss easy items because they interpret the items in a creative manner)
- Random responding (can also happen at the end of a test if the examinee runs out of time)
For assessment we are also interested in:

- Identifying test items that aren’t working well
- Identifying poorly written test items
- Identifying common mistakes from students
- Identify unusual response patterns from students
  - Did the student miss a class where the topic was covered?
  - Was the item biased against students from particular groups?
  - How can I change my teaching to help all students be successful?
Most item analysis techniques:

- Require large samples (at least 300 examinees, maybe more - see Nunnally (1967))
- Can require substantial expertise, computer programming, or use of special software
- Today’s technique will work for small samples (within reason, such as within a single class)
- Uses an Excel template (you can also program it into other systems if you wish)
- Intended for multiple-choice tests (use with rubrics or other types of assessments are a talk for another year!)
- Provides information on the items and on the examinees
Review of Core Concepts

- Item Difficulty
- Item Discrimination
- Distractor Analysis
Item Difficulty
Item Difficulty

- Item difficulty is the proportion of examinees who got an item correct
- Term is a misnomer (really item “easiness”)

\[ P = \frac{\text{Number of correct responses}}{N} \]
Item Difficulty

Item difficulty (p-scores)

1.00 (very easy item)

0.50 (50% answer correct)

0.00 (very hard item)

Examinee ability

θ (typically used)

Low ability

Moderate ability

High ability
Item Difficulty Example

- Everyone got item 2 correct
- Items 4 and 5 were very hard (few got them correct)

<table>
<thead>
<tr>
<th>Item</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.68</td>
</tr>
<tr>
<td>4</td>
<td>0.22</td>
</tr>
<tr>
<td>5</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Jeremy Penn
AALHE, 2017
North Dakota State University
Interpretation depends on the purpose of the test:

**Mastery (criterion-referenced)**
- Purpose is to determine whether or not students gained specific knowledge, skills, abilities
- High item difficulties (items that most students got correct) are acceptable because it suggests students mastered the intended content

**Ranking / sorting (norm-referenced)**
- Purpose is to sort or rank students from “most” to “least” of the thing that you are testing
- High item difficulties (items that most students got correct) are USELESS because they provide no information about which students have “more” or “less” of the thing you are testing

Jeremy Penn
AALHE, 2017
Item Difficulty by Group

- Item 3 may have group bias (differential item functioning, or DIF)
- Check wording of item for possible bias
- Check for equal opportunity to learn
- Could represent actual performance difference between groups

<table>
<thead>
<tr>
<th>Item</th>
<th>Domestic Students</th>
<th>International Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.97</td>
<td>0.39</td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
<td>0.24</td>
</tr>
<tr>
<td>5</td>
<td>0.09</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Jeremy Penn
AALHE, 2017
Item Discrimination: Can you tell the difference?
Item Discrimination

Item discrimination is the extent to which an item differentiates students on the attribute of interest.

That is, the item’s ability to correctly separate those with “high” and “low” knowledge in the subject area or “more” or “less” of whatever you are testing.
Item Discrimination

Maximum discrimination ($D = 1.00$) occurs when everyone in the “high” group gets the item correct and everyone in the “low” group gets the item wrong.

Negative or 0 discrimination is bad!
Calculating Item Discrimination

\[
D = \frac{\text{Number correct (high)}}{N_{\text{high}}} - \frac{\text{Number correct (low)}}{N_{\text{low}}}
\]

- Can split high / low at median if N is small or use top 25% if N is large
- Other methods to calculate discrimination can also be used (e.g., item-total point-biserial correlation)
Item Discriminations

- Generally want discrimination to be 0.30 or higher (but see notes on next slide)
- Item 2 provides no information about which students have “more” or “less” of the thing you are testing
- Item 5 had the biggest gap in performance between the high-achieving group and the low-achieving group

<table>
<thead>
<tr>
<th>Item</th>
<th>Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.30</td>
</tr>
<tr>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
</tr>
<tr>
<td>5</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Interpretation again depends on the purpose of the test:

Mastery (criterion-referenced)
- Purpose is to determine whether or not students gained specific knowledge, skills, abilities
- Will have low discrimination if lots of students get the item correct
- May not be that concerned about low item discriminations

Ranking / sorting (norm-referenced)
- Purpose is to sort or rank students from “most” to “least” of the thing that you are testing
- Would remove items that don’t provide reasonable discrimination
- Why waste time with items that don’t separate students into performance groups?
What do you mean you have to study?
Distractors in a Multiple-Choice Test

• These are the possible response options that are not the correct answer ("distract" from the right answer)

• Analyzing students’ choice of distractor can improve test creation:
  • Distractors that are not chosen at all or infrequently should be replaced (obviously incorrect)
  • Distractors that are selected more than the correct answer might be misleading

• Analyzing students’ choice of distractor can improve teaching:
  • Identify common errors
  • Can be used to modify teaching in the future
Let’s Play a Kahoot to see what you’ve learned!

Each person/team needs an internet-connected device
Modified Caution Index
Developed by Hamisch, Linn, and Sato in the 1980s
Modified Caution Index (MCI)

- Uses item difficulty and item discrimination together to generate a new index
- Higher MCI scores suggest additional “caution” when interpreting score results
- Works reasonably well even on small samples (such as a typical classroom)
- Still requires professional judgment – don’t let item statistics be deterministic
- Easy to calculate in Excel or with statistical software
Key references:


Modified Caution Index: People

• Produces an index from 0 to 1 (≥0.26 generally considered problematic – see Karabatsos, 2003)

• In combination with achievement on the test, the MCI is used to ‘categorize’ the score pattern

\[
C_i^* = \frac{\sum_{j=1}^{n_i} (1 - u_{ij})n_j - \sum_{j=n_i+1}^{J} u_{ij}n_j}{\sum_{j=1}^{n_i} n_j - \sum_{j=J+1-n_i}^{J} n_j}
\]

\(i = 1, 2, \ldots I\), indexes the examinee,
\(j = 1, 2, \ldots J\), indexes the item,
\(u_{ij} = 1\) if examinee \(i\) answers item \(j\) correctly (0 otherwise),
\(n_i\) = total correct for the \(i^{th}\) examinee,
\(n_j\) = total number correct responses to the \(j^{th}\) item
Consider these four students:

A. Got nearly all the items correct except for the very hardest question

B. Got nearly all the items correct but surprisingly missed the two “gimmie” questions

C. Got the five easiest questions correct but missed all the hard questions

D. Got the two hardest questions right but missed everything else
The Modified Caution Index offers the ability to classify these four students:

A - High achieving student with a regular response pattern.
B - High achieving student with an irregular response pattern.
C - Low achieving student with a regular response pattern.
D - Low achieving student with an irregular response pattern.

\[
\begin{array}{c|c}\hline
A & B \\
\hline
C & D \\
\hline
\end{array}
\]

*Can use different values for the MCI and % correct cutoff depending on the context of the test

From: Harnisch and Kelberlau

Jeremy Penn
AALHE, 2017

North Dakota State University
Said another way:

A. Doing everything fine
B. Making careless mistakes (or good at “guessing” the hard items)
C. In need of more study or sporadic habits
D. Insufficient readiness
Modified Caution Index: **Items**

- Same exact calculation as for people, but reverses the role of the people and the items
Consider these four items:

W. Only your best students answered it correctly.
X. Only a few students answered it correctly and it was not your best students.
Y. Most students answered it correctly including all of your best students.
Z. Most students answered it correctly but some of your best students missed it.
The Modified Caution Index has the ability to classify these four items:

W. Difficult item with expected response pattern
X. Difficult item with an irregular response pattern
Y. Easy item with expected response pattern
Z. Easy item with an irregular response pattern

**MCI Code for Items**

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>Z</td>
</tr>
</tbody>
</table>

- **W**: 0.00 MCI
- **X**: 0.30 MCI
- **Y**: 1.00 MCI

Percentage of students answering item correctly:
- 0% (0 students)
- 50% (1 student)
- 100% (2 students)
Said another way:

W. Seems to work well and discriminates between high and low performing students (few students get it correct)

X. Item might need revision, low performing students are getting the item correct while high performing students are missing the item (few students get it correct)

Y. Seems to work well and discriminates between high and low performing students (most students get it correct)

Z. Item might need revision, low performing students are getting the item correct while high performing students are missing the item (most students get it correct)
Item Analysis Using Excel

Jeremy Penn
AALHE, 2017
Item analysis using Excel

• Download the Excel file here: https://goo.gl/pObVkJR
  • Note: this address will be changing later this summer

• Current version limited to 50 items and 159 examinees
  (can be expanded to the limits of Excel)

• Current version does not have support for testlets or subscales
  (analyze separately or use other software)

• Current version does not have support for multiple-group analyses
  (analyze separately or use other software)
Getting Started: Put your data (and answer key) into an Excel file

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Key</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>ID 1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Getting Started: Put your data (and answer key) into an Excel file

People
Getting Started: Put your data (and answer key) into an Excel file

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>q1</td>
<td>q2</td>
<td>q3</td>
<td>q4</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Key:</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>ID 1</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Answer key is the first row
Copy and paste your data into the scoring template on the ‘data’ tab, beginning in cell B2. Then move to tab “Step 1”
Step 1 tab

- Your data have been scored correct / incorrect (1 = correct, 0 = incorrect)
- In order for scoring to be done correctly, YOU MUST DELETE THE ‘1’ OUT OF ALL OF THE CELLS YOU ARE NOT USING
In this example we only had 40 items, so we must delete the ‘1’s out of these columns. DO NOT DELETE THE ENTIRE COLUMN OR ROW, JUST THE DATA.
It will look like this for these 10 items for all 159 respondents.

Move to the “sorted” tab.
You must manually sort the table so the items with the highest difficulty (largest number in row 1) is at the left in cell B1 and the item with the lowest difficulty is all the way to the right.

![Sort Options](image1)

![Sort Result](image2)

---

Jeremy Penn  
AALHE, 2017
Repeat these steps on the “sorted transposed” tab (this will calculate the item indices).
<table>
<thead>
<tr>
<th>ID 18</th>
<th>ID 36</th>
<th>ID 51</th>
<th>ID 115</th>
<th>ID 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Jeremy Penn**  
AALHE, 2017  
North Dakota State University
Move to tab “Step 2”
Verify the information in the green box is correct. You can change the “lowest passing score” percent and the name of the test if you wish.

### Step 2: Check to make sure these are correct.

<table>
<thead>
<tr>
<th>What shall we call this test?</th>
<th>Test 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many items in the test?</td>
<td>40</td>
</tr>
<tr>
<td>How many people took the test?</td>
<td>159</td>
</tr>
<tr>
<td>What is the lowest passing score?</td>
<td>70%</td>
</tr>
</tbody>
</table>

These parameters will be used in our calculations.
**Step 2: Check to make sure these are correct.**

<table>
<thead>
<tr>
<th>Question</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>What shall we call this test?</td>
<td>AALHE Test</td>
</tr>
<tr>
<td>How many items in the test?</td>
<td>40</td>
</tr>
<tr>
<td>How many people took the test?</td>
<td>159</td>
</tr>
<tr>
<td>What is the lowest passing score?</td>
<td>60%</td>
</tr>
</tbody>
</table>

These parameters will be used in our calculations.
Delete the cells for the items you are not using (in this case, #41 – 50). Just delete the data, not the entire column / row.

<table>
<thead>
<tr>
<th></th>
<th>#41</th>
<th>#42</th>
<th>#43</th>
<th>#44</th>
<th>#45</th>
<th>#46</th>
<th>#47</th>
<th>#48</th>
<th>#49</th>
<th>#50</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.54</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.45</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
</tr>
<tr>
<td>0.25</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
</tr>
<tr>
<td>0.6</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
</tr>
<tr>
<td>0.6</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
</tr>
<tr>
<td>(0.0)</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
<td>#DIV/0!</td>
</tr>
<tr>
<td>#40</td>
<td>#41</td>
<td>#42</td>
<td>#43</td>
<td>#44</td>
<td>#45</td>
<td>#46</td>
<td>#47</td>
<td>#48</td>
<td>#49</td>
<td>#50</td>
</tr>
<tr>
<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>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Move to “Step 3”
Delete any cells for examinees you don’t have

• Delete just the cells, not the entire row.
(In our example, we have 159 examinees so there is nothing to delete.)

😊 All that is left is to analyze your results! 😊
Analyzing results
Step 4a tab
### Test Summary Information

<table>
<thead>
<tr>
<th></th>
<th>AALHE Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of items ($k$) in the test</td>
<td>40</td>
</tr>
<tr>
<td>Sample size ($n$)</td>
<td>159</td>
</tr>
<tr>
<td><strong>Average Score ($M$)</strong></td>
<td>28.19</td>
</tr>
<tr>
<td><strong>Cutoff Score</strong></td>
<td>24.00</td>
</tr>
<tr>
<td><strong>Median (midpoint)</strong></td>
<td>29.00</td>
</tr>
<tr>
<td><strong>Lowest Score</strong></td>
<td>5.00</td>
</tr>
<tr>
<td><strong>Highest Score</strong></td>
<td>38.00</td>
</tr>
</tbody>
</table>

70% 60% 73% 13% 95%
Test Reliability

• Three different measures of reliability
• These will be misleading if your test covers different content areas ("Testlets" analyze separately). EG:
  • World history
  • Economics
  • Star Wars Trivia
  • Should be analyzed separately for accurate reliability estimates
Item statistics

- Item #37 has negative discrimination, should be revised
- Item #36 and #15 have borderline discrimination
- #6, #9, #10, #15, #28 all have difficulties outside ideal parameters (but this depends on the purpose of the test, as discussed earlier).
Step 4b tab
Step 4b has a scatter plot that shows the items plotted by discrimination and difficulty.
You can show data labels for these items if you want.

Right click on a data point and select “Add Data Labels,” and “Add Data Labels.”
Then….

Right click again and select “format data labels.”
...then...

Select ‘Value from Cells.’
…then…

Use the “Select Data Label Range” to select the appropriate labels for the items
...then...

Highlight the item names (I typed these in a blank spot in one of the other tabs), and hit enter and ‘Ok.’

In the Format Data Labels you can then unselect “Y Value.”
Positive Discrimination - Difficult
Test items are okay but revise training to provide more support for these skills.

Positive Discrimination - Easy
Consider if items are too easy, possibly due to prior knowledge. Perhaps, these can be dropped from training.

IDEAL RANGE
Positive Item Discrimination and moderate Item Difficulty

Low, Zero, or Negative Discrimination
When a test item exhibits low discriminant validity, use

#28
#29
#30
#31
#32
#33
#34
#35
#36
#37
Step 4c tab
This is a chart for the distribution of scores for examinees - top score of 38, lowest score of 5. Most common score was 31.
Distractors
• For each item the most-selected distractor is highlighted.

• The numbers show the proportion of examinees who selected each distractor.

• For item #1, 0.18 or 18% of respondents selected “2” instead of the correct answer “3.”
For items #36 and #37, most people picked ‘4’. Were they guessing?
Was it a “none of the above” or “all of the above” option?
Was it confusingly-worded?
Or maybe ‘4’ was the right answer and you scored it wrong!
Caution Index: People
Provides the caution index score ("Person MCI") and the category based on cutoff of 0.30 for MCI and 50% for test performance. You can adjust these values in the coding if you wish.

Person ID 2: Scored 19 (48%) but had a regular response pattern (missed the harder items and got the easy items correct).

Person ID 20: Scored 12 (30%) but had an irregular response pattern.

Person ID 13: Scored 34 (85%) but missed some items he/she should have gotten correct.
Caution Index: Items
• This test has many items that are fairly easy
• #36 and #37 identified as needing revision
• #15, for example, is easy (90% got it right) and poorly discriminates (.2)

<table>
<thead>
<tr>
<th></th>
<th>Modified Caution Signal</th>
<th>Item MCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Y-Easy but fair</td>
<td>0.29</td>
</tr>
<tr>
<td>#2</td>
<td>Y-Easy but fair</td>
<td>0.28</td>
</tr>
<tr>
<td>#3</td>
<td>Z-Easy and needs revision</td>
<td>0.32</td>
</tr>
<tr>
<td>#4</td>
<td>Y-Easy but fair</td>
<td>0.20</td>
</tr>
<tr>
<td>#5</td>
<td>Y-Easy but fair</td>
<td>0.28</td>
</tr>
<tr>
<td>#6</td>
<td>Y-Easy but fair</td>
<td>0.24</td>
</tr>
<tr>
<td>#7</td>
<td>Z-Easy and needs revision</td>
<td>0.38</td>
</tr>
<tr>
<td>#8</td>
<td>Y-Easy but fair</td>
<td>0.19</td>
</tr>
<tr>
<td>#9</td>
<td>Y-Easy but fair</td>
<td>0.31</td>
</tr>
<tr>
<td>#10</td>
<td>Y-Easy but fair</td>
<td>0.20</td>
</tr>
<tr>
<td>#11</td>
<td>Y-Easy but fair</td>
<td>0.24</td>
</tr>
<tr>
<td>#12</td>
<td>Z-Easy and needs revision</td>
<td>0.35</td>
</tr>
<tr>
<td>#13</td>
<td>Y-Easy but fair</td>
<td>0.22</td>
</tr>
<tr>
<td>#14</td>
<td>Y-Easy but fair</td>
<td>0.29</td>
</tr>
<tr>
<td>#15</td>
<td>Z-Easy and needs revision</td>
<td>0.41</td>
</tr>
<tr>
<td>#16</td>
<td>Z-Easy and needs revision</td>
<td>0.35</td>
</tr>
<tr>
<td>#17</td>
<td>Z-Easy and needs revision</td>
<td>0.32</td>
</tr>
<tr>
<td>#18</td>
<td>Y-Easy but fair</td>
<td>0.26</td>
</tr>
<tr>
<td>#19</td>
<td>Y-Easy but fair</td>
<td>0.16</td>
</tr>
<tr>
<td>#20</td>
<td>Y-Easy but fair</td>
<td>0.31</td>
</tr>
<tr>
<td>#21</td>
<td>Y-Easy but fair</td>
<td>0.16</td>
</tr>
<tr>
<td>#22</td>
<td>Z-Easy and needs revision</td>
<td>0.36</td>
</tr>
<tr>
<td>#23</td>
<td>Y-Easy but fair</td>
<td>0.29</td>
</tr>
<tr>
<td>#24</td>
<td>Y-Easy but fair</td>
<td>0.29</td>
</tr>
<tr>
<td>#25</td>
<td>Y-Easy but fair</td>
<td>0.20</td>
</tr>
<tr>
<td>#26</td>
<td>Y-Easy but fair</td>
<td>0.18</td>
</tr>
<tr>
<td>#27</td>
<td>Y-Easy but fair</td>
<td>0.25</td>
</tr>
<tr>
<td>#28</td>
<td>Z-Easy and needs revision</td>
<td>0.33</td>
</tr>
<tr>
<td>#29</td>
<td>W-Difficult but fair</td>
<td>0.20</td>
</tr>
<tr>
<td>#30</td>
<td>Y-Easy but fair</td>
<td>0.22</td>
</tr>
<tr>
<td>#31</td>
<td>Y-Easy but fair</td>
<td>0.17</td>
</tr>
<tr>
<td>#32</td>
<td>Z-Easy and needs revision</td>
<td>0.36</td>
</tr>
<tr>
<td>#33</td>
<td>Z-Easy and needs revision</td>
<td>0.32</td>
</tr>
<tr>
<td>#34</td>
<td>Y-Easy but fair</td>
<td>0.31</td>
</tr>
<tr>
<td>#35</td>
<td>Y-Easy but fair</td>
<td>0.21</td>
</tr>
<tr>
<td>#36</td>
<td>Y-Easy but fair</td>
<td>0.27</td>
</tr>
<tr>
<td>#37</td>
<td>Y-Easy but fair</td>
<td>0.27</td>
</tr>
</tbody>
</table>
What does all this mean?
If all you do when you complete an assessment is summarize the overall scores, you are missing out on a tremendous amount of important information!

Jeremy Penn
AALHE, 2017
North Dakota State University
Item analysis provides information about examinees

• Teach your students how to be good students!
• Use this information to work with students who made careless mistakes – if they can correct the mistakes, they will likely greatly improve!
• Those who need more study have the basics but struggle with the more difficult tasks – remediate and get caught up
Will item analysis prove a student has cheated?

• No, but it can be one component of building a case when cheating is suspected
Item analysis provides information about the test and your teaching

- Based on personal experience alone, most classroom tests are not very good
- Careful attention to the performance of the items in your exams can help you improve your tests – remove misleading or biased items
- Helps you be clear about the purpose of testing
Look for biased items

Examine the items coded as X ("difficult and needs revision") and Z ("easy and needs revision")
Can statistically test for group differences on individual items

<table>
<thead>
<tr>
<th>Item #36</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td>5</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Group B</td>
<td>8</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>20</td>
<td>33</td>
</tr>
</tbody>
</table>

Statistically significant in this example, $p = 0.04$. 25% of group A got it correct vs. 62% of group B.

Online calculation of statistics for 2x2 contingency tables: http://vassarstats.net/odds2x2.html

(Should have expected frequency of at least 5 for each cell.)
Interview students!

• Group differences do not prove the existence of bias. Sometimes groups score differently on items and this can represent a real difference.

• You can simply ask students to explain their reasoning and how they answered the item – this can reveal more about how students were thinking about the item and whether or not it is biased.
Can statistics of this sort be automatically programmed into the testing software your institution uses?
Use item analysis as a tool to improve student learning! Isn’t that why we are all here?