KEY VALIDATION: HOW TO LOCATE POORLY FUNCTIONING ITEMS TO IMPROVE EXAM QUALITY

ANDREW DALLAS, FEN FAN, J. B. WEIR




Paving the way for future PAs

OVERVIEW OF WORKSHOP

- The importance of assessment
- Characteristics of quality assessments
- Finding Bad Items: Item analysis & statistics
- Flagging Items: Identifying flawed items using statistical criteria
- Key Validation: Using statistics in your content review

Click to add text

NCCPA

WHY DO WE ASSESS?

- Goals of assessment in the classroom
- Informal vs. formal assessments
- Summative or formative

NCCPA

QUALITY ASSESSMENTS

An instrument that produces *valid* scores…

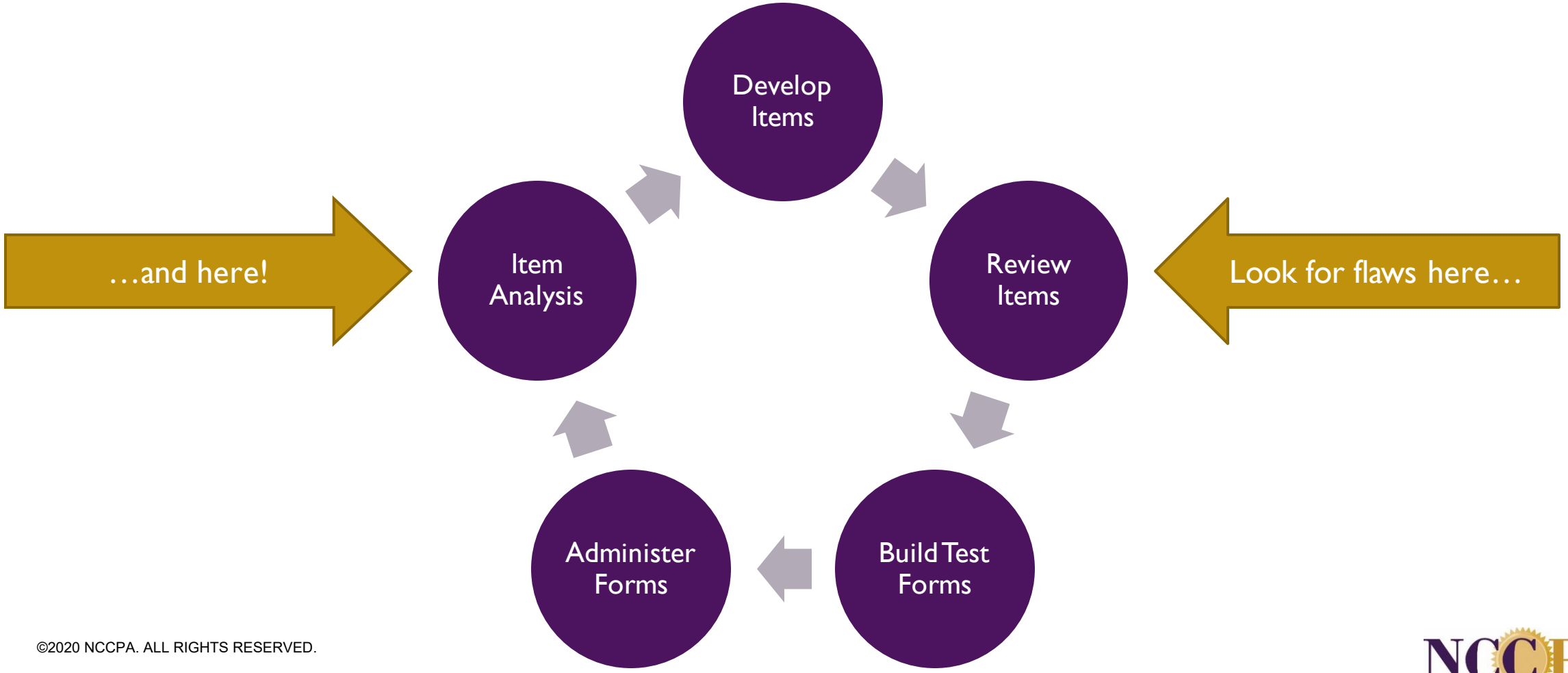
- measures what is intended,
- can be used to make meaningful decisions, and
- contains content that is current and up-to-date.

An instrument that produces *reliable* scores provides a consistent picture.

NCCPA

FINDING BAD ITEMS: ITEM ANALYSIS & STATISTICS

A TYPICAL EXAM CYCLE



NCCPA

ITEM ANALYSIS

- Item analysis is a statistical analysis of the response data gathered from an assessment.
- This type of analysis provides clues about which items are working well and which items may need a second look.
- Today, we will focus on just a few item-evaluation metrics.

NCCPA

ITEM STATISTICS

- The proportion of people who answer an item correctly (the *p*-value)
- The proportion of people selecting each incorrect option (distractor proportions)
- The relationship between performance of a specific item and the performance of the total exam (point-biserial)
- We use these three measures in combination to help find flawed items

NCCPA

ITEM STATISTICS

What might constitute evidence that an item is functioning properly?

- The proportion of people answering correctly should be greater than chance.
- The key is selected most often.
- The probability of an examinee getting the question correct increases with a test-taker's ability (positive R).

NCCPA

ITEM STATISTICS

What might constitute evidence that an item is ***not*** functioning properly?

- P value is below chance.
- Negative, or near-zero, point-biserial (low performers are more or as likely to answer correctly).
- One or more answer option is selected at a high rate.

NCCPA

A SHORT QUIZ

An item has a p-value of .75 and a point-biserial of .20

- Is this a good item?
- How do you know?

NCCPA

A SHORT QUIZ

An item has a point-biserial near zero and a *p*-value of 0.15.

- Would you want to take a second look?
- What other statistics might help determine if there is an issue with this item?

NCCPA

A SHORT QUIZ

An item has a *p*-value of 0.95.

- Is this item okay?
- What other statistics would you want to look to help you decide?

NCCPA

EXAMPLE OF ITEM ANALYSIS OUTPUT

Do the *p* and point-biserial values give us any pause?

Content Description	Content Value
Exam	Mid-Term Exam
Content	Endocrine
Diagnosis	Hyperparathyroidism
P-Value	0.45
Point Biserial	-0.06

Item Statistics	A	B	C	D	E
Point Biserial	0.11	-0.06	0.18	-0.13	-0.02
Proportion (all)	0.03	0.45	0.18	0.21	0.13
Proportion-Lowest Performers	0.00	0.53	0.13	0.20	0.13
Proportion-Average Performers	0.08	0.33	0.12	0.29	0.17
Proportion-Highest Performers	0.00	0.46	0.38	0.08	0.08

NCCPA

EXAMPLE OF ITEM ANALYSIS OUTPUT

How about these *p* and point-biserial values?

What is the correct answer?

The highest performers give us a clue.

Content Description	Content Value
Exam	Mid-Term Exam
Content	Endocrine
Diagnosis	Hyperparathyroidism
P-Value	0.63
Point Biserial	0.29

Item Statistics	A	B	C	D	E
Point Biserial	0.29	-0.20	-0.05	-0.15	-0.15
Proportion (all)	0.63	0.10	0.18	0.05	0.04
Proportion-Lowest Performers	0.44	0.19	0.21	0.08	0.08
Proportion-Average Performers	0.65	0.09	0.20	0.04	0.03
Proportion-Highest Performers	0.84	0.03	0.12	0.01	0.01

This kind of stair-stepping is what we like to see.

NCCPA

CONDUCTING AN ANALYSIS

- An item analysis can be complicated and typically requires special software.
- However, for small-scale analyses, most item statistics can be computed in an EXCEL spreadsheet using three functions:
  - AVERAGE
  - CORREL
  - COUNTIF

NCCPA

QUESTIONS ON ITEM STATISTICS?

- Do you look at any item statistics as you evaluate your own test instruments?
- Are there other elements of items that we might examine?

NCCPA

FLAGGING ITEMS

FLAGGING ITEMS

- The purpose of the item analysis is to identify items that are in need of additional content review.
- Establishing appropriate statistical criteria for reviewing items is essential.
- The next activity will focus on how you might go about flagging items for review.

NCCPA

FLAGGING ITEMS

- Different flags may make sense in different scenarios.
- Typical reasons for flagging an item may include:
  - Too easy (Are we giving the item away with a clue in the stem?)
  - Too hard (Is this miskeyed, not appropriate for the population, or have the standards of care changed?)
  - Poor/Negative discrimination (Is this miskeyed? Is there another answer that could be correct?)

NCCPA

FLAGGING ITEMS

- Flags are just an indication that something could be wrong.
  - Item statistics are just a way to start the process.
  - After flagging an item, a content review should be conducted.
- All flags are not equally critical.
- Sometimes it's key to take your population into account when judging your statistics.

NCCPA

QUESTIONS ON FLAGGING?

- Have you ever taken an exam and thought, "Something's wrong with this question; it shouldn't be on here"?
- Is it okay to have a very easy item on an exam? (Is it okay to have a very low point-biserial?)
- Do you have questions?

NCCPA

KEY VALIDATION

REVIEWING FLAGGED ITEMS

- Key validation is the process we use to review items that have been flagged as a result of item analysis.
- The process requires at least one content expert, but ideally more, to review items.

NCCPA

REVIEWING FLAGGED ITEMS

- A committee of subject-matter experts reviews each item focusing on the content while keeping in mind:
  - Certain statistical clues can be used to help point to potential problems.
  - Items will be flagged even though there is no content-based rationale—we should not rely solely on the statistics for item removal.
- Remember the purpose of the exercise is to identify flawed items, not to critique the item writer, or rewrite the items during key validation.

NCCPA

MAKING DECISIONS

- Keep
  - Item is current and clinically relevant
  - No flaws are identified in the review
- Rewrite
  - Item has flaws
  - Assesses appropriate content
  - Can salvaged with minimal changes
- Delete
  - item is not current or clinically relevant
  - Major effort to correct identified flaws

NCCPA

QUESTIONS ON KEY VALIDATION?

- What might a key validation look like in your context?
- Who would you convene for a key validation?
- Do you have questions?

THANK YOU