Are Your Multiple-Choice Tests “FIT”? 
Using the Fairness of Items Tool (FIT) as a Component of the Test Development Process

Nikole Anderson Hicks, PhD, RNC, CNE
Background

- Multiple-choice (MC) examinations are common
- Conclusions based on MC examinations have high stakes consequences
- Developing well-constructed test items is difficult & time consuming
- Nurse educators lack adequate preparation & sufficient time
- Use of textbook test bank items in examinations is common
- Violations of item-writing guidelines in nursing examinations are common
Theoretical Foundations

• High quality test items are necessary for reliable, valid, discriminating, and unbiased assessments of student learning.

• Item quality is improved through item-writing procedures, obtaining pretest reliability data, and using post-administration analysis data to guide revision.

• Test quality is improved through adequate planning of assessments & developing a test blueprint.
Framework for Quality Assessment

Every effective assessment must meet the following criteria:

- **Valid** – measures what it is designed to measure
- **Reliable** – consistently measures what it is designed to measure
- **Discriminating** – distinguishes between the more knowledgeable & less knowledgeable students
- **Practical** – useful & practical for its purposes
- **Unbiased** – fair to examinees & contains items that students of equal ability are equally likely to answer correctly
Framework for Test Development

**Blueprint**
- School philosophy
- Curricular outcomes
- Nursing course objectives
- NCLEX test blueprints

**Exam Creation Phase**
- Establish reason for the test
- Determine best type of test items
- Develop a test blueprint
- Determine scoring criteria
- Sequence test items
- Define administration procedures

**Assumption of Classical Test Theory**
- Measurement of educational attributes is useful in predicting performance

**Expert Clinicians**
- Nursing faculty writing test items:
  - are clinical experts
  - are proficient in item writing practices

**Test Item Writing Phase**
- Include rationale for each test item
- Write questions at the application or above cognitive level
- Require multi-logical thinking to answer questions
- Require high level of discrimination to choose from among plausible alternatives
- Include identification and revision of biased items

**Validity**
- As defined by classical test theory
  - Content
  - Construct
  - Criterion-related

**Exam Evaluation Phase**
- Include identification and revision of biased items

**Bloom's Taxonomy**

**Paul's Critical Thinking Theory**

**Validity**
- Graduate follow-up study
- NCLEX examination outcomes

**Reliability**
- Assess test items
  - Difficulty level
  - Item discrimination (PBCC)
  - Reliability coefficient (KR20)
Review of Literature

- Interventions that improve item writing:
  - Faculty education & practice
  - Using pre-established guidelines
  - Peer review process
  - Revising items through linguistic modification
Methodology

• Phase 1 – Developing the Fairness of Items Tool (FIT)
• Phase 2 – Expert Review & Validation
• Phase 3 – Use of the FIT by Nursing Faculty

This research study received approval from the Institutional Review Boards at University of Northern Colorado and the University of Cincinnati.
## Phase 2 Results

<table>
<thead>
<tr>
<th>Validity Index</th>
<th>Review 1</th>
<th>Review 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-CVI</td>
<td>.90</td>
<td>.988</td>
</tr>
<tr>
<td>S-CVI/UA</td>
<td>.63</td>
<td>.97</td>
</tr>
<tr>
<td>Face Validity</td>
<td>1.0</td>
<td>.92</td>
</tr>
<tr>
<td>Proportion Relevant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert 1 = .93</td>
<td></td>
<td>Expert 1 = 1.0</td>
</tr>
<tr>
<td>Expert 2 = .98</td>
<td></td>
<td>Expert 2 = 1.0</td>
</tr>
<tr>
<td>Expert 3 = .73</td>
<td></td>
<td>Expert 3 = .97</td>
</tr>
<tr>
<td>Expert 4 = .90</td>
<td></td>
<td>Expert 4 = .97</td>
</tr>
<tr>
<td>Expert 5 = .98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACP</td>
<td>.90</td>
<td>.99</td>
</tr>
</tbody>
</table>

**Notes:**
- **S-CVI**: Scale item content validity index
- **S-CVI/UA**: Universal calculation method for the scale item content validity index
- **ACP**: Average congruency percentage
Fairness of Items Tool (FIT)

Evaluate the Stem
1. Use a question format.
2. Eliminate extraneous words (e.g., *of the following*).
3. Present a single, clearly defined question with the problem in the stem.
4. Avoid negatively phrased questions, double negatives, and the use of *except*.
5. Use active verbs and present tense.
6. Write questions at the application or above cognitive level.
7. Write questions that require multi-faceted thinking (require knowledge of more than one fact/concept).
8. Make sure content is current.
9. Avoid testing student opinions (e.g., use *nurse* instead of *you* as the subject).
10. Test important content and avoid trivia.

Evaluate the Options
11. Make sure options are similar grammatically and in length and amount of detail.
13. Avoid negatively phrased options.
14. Avoid repeating material in the options – move repetitive words to the stem.
15. Avoid repeating words in the stem and correct option.
16. Avoid overlapping options.
17. Eliminate multiple-multiples.
18. Make sure all distractors are plausible.
19. If the stem asks what should be done first or which action is best, all options must be correct with only one option being the first or best.
20. Make sure there is only one correct answer.
21. Write options that require a high level of discrimination to select the correct answer.

Linguistic/Structural Bias
22. Use a parsimonious style and short simple sentences.
23. Use correct grammar, punctuation, capitalization, and spelling.
24. Use precise terms (avoid frequently, appropriate).
25. Avoid absolute terms (always, never, all).
26. Use straightforward, uncomplicated language. Test nursing content, not vocabulary or reading.
27. Write items that can be comprehended on the first reading. Avoid tricky or misleading items.
28. Ensure that items are independent of each other.
29. Be specific and clear with directions.
30. Use consistent spacing, question numbering/lettering, page numbering. Make sure options appear on the same page as the question.

Cultural Bias
31. Avoid dominant culture (literature, music, movies, sports, foods) unless essential to safe, effective nursing practice.
32. Eliminate all names.
33. Eliminate all slang.
34. Use terminology from textbook, notes, and common words (home vs. abode).
35. Eliminate humor.
36. Avoid stereotyping and over-representation of cultural groups.
37. Use gender-specific language only when necessary to test nursing content.
38. Present the person first, not the diagnosis.
Phase 3 Results

Equivalence – Independence of Scores

• 1,190 values tested ($p < .05$)
• 95.5% demonstrated independence ($n = 1,136$)

Equivalence – Interrater Agreement

<table>
<thead>
<tr>
<th>Interpretation of Agreement</th>
<th>$n$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect Agreement (.9-.1.0)</td>
<td>240 (46.8%)</td>
</tr>
<tr>
<td>Excellent Agreement (.8-.89)</td>
<td>100 (19.5%)</td>
</tr>
<tr>
<td>Very Good Agreement (.7-.79)</td>
<td>76 (14.8%)</td>
</tr>
<tr>
<td>Good Agreement (.6-.69)</td>
<td>47 (9.2%)</td>
</tr>
<tr>
<td>Fair Agreement (.5-.59)</td>
<td>48 (9.4%)</td>
</tr>
<tr>
<td>Poor Agreement (Below .5)</td>
<td>2 (0.4%)</td>
</tr>
</tbody>
</table>

Stability – Split-half Reliability

• KR-20 = .799 ($\alpha = .05$)
Phase 3 Results

Construct Validity – Known Groups Comparison

<table>
<thead>
<tr>
<th>Level</th>
<th>Known Biased Item(s)</th>
<th>Known Fair Item(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guideline</td>
<td>2.7 +/- 2.4</td>
<td>0.92 +/- 1.4</td>
</tr>
<tr>
<td>Dimension (ST)</td>
<td>0.29 +/- 0.41</td>
<td>0.1 +/- 0.13</td>
</tr>
<tr>
<td>Test Item</td>
<td>7.84 +/- 4.8</td>
<td>2.76 +/- 2.84</td>
</tr>
</tbody>
</table>

Internal Consistency Reliability – Cronbach’s alpha

<table>
<thead>
<tr>
<th>Test Item</th>
<th>$\alpha$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-18</td>
<td>.737</td>
<td>67</td>
</tr>
<tr>
<td>B-13</td>
<td>.73</td>
<td>70</td>
</tr>
<tr>
<td>B-1</td>
<td>.706</td>
<td>66</td>
</tr>
<tr>
<td>B-11</td>
<td>.694</td>
<td>80</td>
</tr>
<tr>
<td>B-35</td>
<td>.651</td>
<td>65</td>
</tr>
<tr>
<td>F-10</td>
<td>.598</td>
<td>67</td>
</tr>
</tbody>
</table>
Recommendations

• Use a systematic process for test development that incorporates the FIT for developing high quality MC test items.

• Use the FIT to develop item banks of quality MC test items to save time in test development.

• Incorporate the FIT in faculty development. Repetition & practice will lead to improvement in MC test items.

• Develop a test review process with faculty peer review using the FIT.

• Three-option items should be implemented as a standard alternative in nursing education.
Implications

• Improving the quality of MC test items used in nursing examinations has the potential to improve student success and better prepare all nursing students for licensure and certification examinations.

• Indirectly, the FIT has the potential to increase the quality, quantity, and diversity of nurses joining the workforce. These improvements in student success also have a positive impact on nursing program accreditation rates and ability to recruit high quality students.

• Improving student success benefits faculty with improved evaluations and less time devoted toremediating students who are performing poorly on examinations containing biased test items.
Acknowledgements

• This research study used Research Electronic Database Capture (REDCap), a secure, web-based survey tool and database, supported by Center for Clinical and Translational Science and Training grant UL1-RR026314

• Thank you to the following individuals who contributed to the development of this research study:
  • Dr. Janice Hayes, Dr. Faye Hummel, Dr. Vicki Wilson, and Dr. Lisa Rue ~ University of Northern Colorado
  • Dr. Jun Ying ~ University of Cincinnati
  • Dr. Kathleen LaSala ~ University of South Carolina
References


