NURSE EDUCATORS’ PRACTICES IN THE MEASUREMENT OF STUDENT ACHIEVEMENT USING MULTIPLE-CHOICE TESTS

Susan Birkhead, DNS, MPH, RN, CNE
Linnea Jatulis, PhD, RN
April 2016
# Faculty Disclosure

<table>
<thead>
<tr>
<th>Faculty name:</th>
<th>Susan Birkhead, DNS, MPH, RN, CNE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conflicts of Interest:</td>
<td>None</td>
</tr>
<tr>
<td>Employer:</td>
<td>Samaritan Hospital School of Nursing, Troy NY</td>
</tr>
<tr>
<td>Sponsorship/Commercial</td>
<td>None</td>
</tr>
<tr>
<td>Support:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Faculty name:</th>
<th>Linnea Jatulis, PhD, RN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conflicts of Interest:</td>
<td>None</td>
</tr>
<tr>
<td>Employer:</td>
<td>The Sage Colleges, Troy NY</td>
</tr>
<tr>
<td>Sponsorship/Commercial</td>
<td>None</td>
</tr>
<tr>
<td>Support:</td>
<td></td>
</tr>
</tbody>
</table>
Goals and Objectives

- **Session Goal:**
  - To identify the importance of adherence to best practices in test construction, item writing and analysis of test results.

- **Session Objectives:**
  
  The learner will be able to:
  
  - list seven best practices in item writing, test construction and analysis of test results.
  
  - contrast and compare faculty practices in assessment of student achievement with respect to faculty educational preparation, teaching experience and program type (AD or BS).
  
  - identify the two most important sources cited by faculty for learning about how to write test questions, how to construct tests, and how to analyze test results.
Introduction

- MCQ tests commonly accepted practice in nursing

- MCQ tests impact student progression

  - False positives: poorly prepared students progress

  - False negatives: knowledgeable students are prevented from progressing
Therefore...

- Tests must be reliable
- Results must be supported by strong validity evidence

- FACULTY COMPETENCE IS ESSENTIAL
Pedagogical Skill

Faculty incompetence “is hard to substantiate on the basis of published evidence, but what is indisputable is that many tests are indeed badly written and administered” (Burton, 2005, p. 66).
Faculty must be able to...

- Write educational objectives according to a cognitive taxonomy
- Construct tables of specification (“blueprints”)
- Write meaningful items
- Analyze test results
Institutional Policies and Practices that Support Excellence

- Test committee
- Test policy
- Test blueprinting
- Test vetting
- Test analysis software
Purpose of the Study

The purpose of this study was to explore nurse educators’ practices in the measurement of student achievement using testing in prelicensure nursing education in New York State.
Purposes of Testing - NLN

- To evaluate student achievement
- To support student learning
- To improve teaching
- To guide program improvement
Literature Review
Guidelines

- Standards for Educational and Psychological Testing (AERA, APA, NCME)
- Code of Fair Testing Practices in Education
- NLN Fair Testing Guidelines
- Code of Ethics for Nurse Educators (Rosenkotter & Milstead)
Taxonomy of MCQ Item-writing Guidelines

- Content concerns
- Formatting concerns
- Style concerns
- Writing the stem
- Writing the choices

Haladyna, Downing & Rodriguez (2002)
Research

- Masters et al. (2001): 2913 MCQs from 17 test banks accompanying nursing textbooks
  - 2233 contained flaws
  - 72% knowledge/comprehension level

- Bosher & Bowles (2008)
  - Linguistic modification adds to the validity evidence supporting the measurements obtained by MCQ tests
Research

• Killingsworth, Kimble, & Sudia (2015):
  – Reported high adherence to best practices
  – Belief that rules are important increased use of best practices
  – Baccalaureate programs only

• Institutional Practices in Exam Construction
  • No studies in nursing
Methodology
Research Design

- Exploratory
- Descriptive
- Correlational
- Quantitative
- Survey
Population

• Anonymous
• Nurse educators, AD or BSN programs in NYS, regardless of status (FT, PT, tenured, etc.) or degree credentials
• Must have classroom teaching responsibilities, including full or shared responsibilities for assessing student achievement using written tests
Independent Variables

- Educational preparation
- Years in nursing education
- Type of nursing education program (AD, BS, accelerated)
- Program accreditation status
- Program size (number of students)
- Teaching modality (independent or team)
- Descriptive institutional factors
Best Practice Index (BPI): dependent variables

- Blueprinting
- Writing rationales
- Writing three total answer choices
- Avoiding none-of-the-above (NOTA)
- Avoiding all-of-the-above (AOTA)
- Using readability analyses
Best Practice Index
(dependent variables)

- Editing commercially prepared test bank questions prior to use
- Vetting test questions
- Using e-scoring system
- Nullifying
- Using the correct nullification procedure
- Debriefing
Results
Summary Demographics

- Response rate: 297/1559 = 19%
- 80% age 50 or older
- 33% had been teaching more than 20 years
- 45% in AD programs, 56% in BSN programs
- 51% education major, 49% other major
- 57% independent, 43% team teach
Most Important Source of Learning for Test Construction

- Professional literature: 3%
- Training (NCSBN, etc.): 8%
- Other: 8%
- Formal coursework: 21%
- Mentoring: 29%
- CE Conferences: 31%
Analysis
Research Question One

What is the prevalence of the use of multiple-choice testing to measure student achievement in prelicensure nursing education in New York State?
Reported Use of Question Types

Mean percent question type reported

Question type

- Essay
- True-false
- Matching
- Short answer
- Fill-in for dose calculation
- Other alternate format
- Multiple response
- Standard MCQs

The graph shows the percentage of each question type reported, with Standard MCQs being the most commonly used.
Use of Standard Four-Answer MCQs

Percent of test questions that are standard four-answer MCQs

Number of respondents

0 - 19%: 1%
20 - 39%: 3%
40 - 59%: 3%
60 - 79%: 17%
80 - 99%: 69%
100%: 6%
### Extent of MCQ Use by Program Type

<table>
<thead>
<tr>
<th>Percent of questions on a typical test that are standard four-answer MCQs</th>
<th>0 -19%</th>
<th>20 - 39%</th>
<th>40 - 59%</th>
<th>60 - 79%</th>
<th>80 - 99%</th>
<th>100%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = Associate degree</strong></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>14</td>
<td>65</td>
<td>2</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>0.0%</td>
<td>1.2%</td>
<td>2.4%</td>
<td>16.7%</td>
<td>77.4%</td>
<td>2.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>n = Baccalaureate degree</strong></td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>20</td>
<td>60</td>
<td>9</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>1.0%</td>
<td>5.1%</td>
<td>4.0%</td>
<td>20.2%</td>
<td>60.6%</td>
<td>9.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>34</td>
<td>125</td>
<td>11</td>
<td>183</td>
</tr>
<tr>
<td></td>
<td>0.5%</td>
<td>3.3%</td>
<td>3.3%</td>
<td>18.6%</td>
<td>68.3%</td>
<td>6.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

\( \chi^2(5) = 8.877, p > .05 \)
Proportion of Course Grade Derived from Tests

- <20%: 10 respondents
- 20 - 39%: 20 respondents
- 40 - 59%: 30 respondents
- 60 - 79%: 40 respondents
- 80 - 99%: 60 respondents
- 100%: 70 respondents
Ninety-nine respondents (50%) reported that at least 80% of the course grade was based on tests that were composed of at least 80% standard four-answer multiple-choice questions.
Course Grades from MCQs

Percent of course grade derived from tests

Percent of test questions that are standard four-answer MCQs
# Proportion of Course Grade Derived from Tests by Program Type

<table>
<thead>
<tr>
<th>Proportion of course grade derived from tests</th>
<th>&lt;20%</th>
<th>20 - 39%</th>
<th>40 - 59%</th>
<th>60 - 79%</th>
<th>80 - 99%</th>
<th>100%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = Associate degree</strong></td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>39</td>
<td>38</td>
<td>89</td>
</tr>
<tr>
<td><strong>Proportion</strong></td>
<td>5.6%</td>
<td>1.1%</td>
<td>0.0%</td>
<td>6.7%</td>
<td>43.8%</td>
<td>42.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>n = Baccalaureate degree</strong></td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>30</td>
<td>39</td>
<td>14</td>
<td>111</td>
</tr>
<tr>
<td><strong>Proportion</strong></td>
<td>8.1%</td>
<td>8.1%</td>
<td>9.0%</td>
<td>27.0%</td>
<td>35.1%</td>
<td>12.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>14</td>
<td>10</td>
<td>10</td>
<td>36</td>
<td>78</td>
<td>52</td>
<td>200</td>
</tr>
<tr>
<td><strong>Proportion</strong></td>
<td>7.0%</td>
<td>5.0%</td>
<td>5.0%</td>
<td>18.0%</td>
<td>39.0%</td>
<td>26.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

χ²(5) = 42.717, p < .05
Research Question Two

What is the relationship between nurse educator characteristics (educational preparation, age, years teaching nursing) and self-reported practices in the use of testing to measure student achievement?
<table>
<thead>
<tr>
<th>Major</th>
<th>Best Practice Index Score</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;5</td>
<td>5-8</td>
<td>≥9</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Other major</td>
<td>9</td>
<td>52</td>
<td>37</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9%</td>
<td>53%</td>
<td>38%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Education major</td>
<td>11</td>
<td>62</td>
<td>29</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11%</td>
<td>61%</td>
<td>28%</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>114</td>
<td>66</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>57%</td>
<td>33%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

χ²(2) = 1.968, p > .05
Additional Findings

- Age and years teaching were not significant predictors of BPI scores.

- Those who taught in teams were significantly more likely to have their tests vetted ($p < .05$).

- Thirty-four percent ($n = 63$) never nullify questions.
Use of Reliability Data

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>17%</td>
</tr>
<tr>
<td>Yes</td>
<td>54%</td>
</tr>
<tr>
<td>I am not sure what the KR-20 is</td>
<td>25%</td>
</tr>
</tbody>
</table>
Research Question Three

What is the relationship between nursing education program institutional characteristics (program type, program size, accreditation status, policies) and nurse educators’ self-reported practices in the use of testing to measure student achievement?
# Respondents’ BPI Scores by Program Type

<table>
<thead>
<tr>
<th>Program type</th>
<th>Best Practice Index score</th>
<th>Less than 5</th>
<th>5 - 8</th>
<th>9 or more</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td></td>
<td>8</td>
<td>53</td>
<td>28</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.0%</td>
<td>59.6%</td>
<td>31.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>BS</td>
<td></td>
<td>12</td>
<td>61</td>
<td>38</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.8%</td>
<td>55.0%</td>
<td>34.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>20</td>
<td>114</td>
<td>66</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.0%</td>
<td>57.0%</td>
<td>33.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

$k^2(2) = .462, p > .05$
## Existence of a Testing Policy by Program Type

<table>
<thead>
<tr>
<th></th>
<th>Testing policy</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Total</td>
</tr>
<tr>
<td>AD</td>
<td>21</td>
<td>63</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>25.0%</td>
<td>75.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>BS</td>
<td>51</td>
<td>52</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>49.5%</td>
<td>50.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>115</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td>38.5%</td>
<td>61.5%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

$\chi^2(1) = 11.743, p < .05$
## Existence of a Policy Requiring Blueprinting by Program Type

<table>
<thead>
<tr>
<th>Program has a policy requiring blueprinting</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>39</td>
<td>45</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>46.4%</td>
<td>53.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td>BS</td>
<td>94</td>
<td>9</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>91.3%</td>
<td>8.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>133</td>
<td>54</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td>71.1%</td>
<td>28.9%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

$\chi^2(1) = 45.281, \ p < .05$
The Effect of a Policy Requiring Blueprinting on the Use of Blueprinting

<table>
<thead>
<tr>
<th>Reported frequency of use of test blueprint</th>
<th>Never</th>
<th>Occasionally</th>
<th>Often</th>
<th>Always</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No policy requiring blueprinting</td>
<td>14</td>
<td>32</td>
<td>38</td>
<td>49</td>
<td>133</td>
</tr>
<tr>
<td>Policy requiring blueprinting</td>
<td>1</td>
<td>7</td>
<td>12</td>
<td>34</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>39</td>
<td>50</td>
<td>83</td>
<td>187</td>
</tr>
</tbody>
</table>

$\chi^2(3) = 12.354, p < .05$
## Existence of a Policy Requiring Vetting by Program Type

<table>
<thead>
<tr>
<th>Program Type</th>
<th>Program has a policy requiring vetting</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>AD</td>
<td>55</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>65.5%</td>
<td>34.5%</td>
</tr>
<tr>
<td>BS</td>
<td>86</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>83.5%</td>
<td>16.5%</td>
</tr>
<tr>
<td>Total</td>
<td>141</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>75.4%</td>
<td>24.6%</td>
</tr>
</tbody>
</table>

$\chi^2(1) = 8.099, \ p < .05$
Conclusions

- Testing is heavily used in prelicensure nursing education programs
- MCQ tests predominate
- A large part of course grades are derived from results of MCQ tests
- Little use of alternate format questions
Conclusions

No significant relationship between respondent

- Age
- Years teaching
- Educational preparation

... and scores on the Best Practice Index.

• Homogeneity
Conclusions

- Continuing education and mentoring most important sources of learning.

- No significant relationship between program type and respondents’ BPI scores.

- AD programs were more likely to have written testing policies.

- Respondents were significantly more likely to report engaging in blueprinting and vetting when required to do so by policy.
Additional Conclusion

NYS nurse educators are ... MATURE.
Limitations

- Geographic limitations
- Low response rate
- Not all NYS programs publish faculty emails online
- Selection bias
- Potential reporting bias (positive adherence)
- Descriptive, cross-sectional survey: correlations do not imply cause and effect relationships
- Implication only: adherence to best practices results in better tests
Recommendations

• Graduate programs for nursing education majors: require formal coursework in assessment of student achievement.

• Professional organizations: frequent, accessible continuing education programs.

• Accrediting bodies:
  – establish standards
  – closely examine education programs’ practices.
Recommendations

- Novice nurse educators: NCSBN Learning Extension online courses.
- Mentors assigned to novice educators.
- Workloads adjusted for time to construct impeccable assessments.
- Educators should view assessment as a scholarly activity, reflective of teaching abilities.
Areas for Further Research

- Existence and nature of testing policies
- Use of blueprinting
- Vetting practices
- Test debriefing practices
- Reproduce the study by Tarrant, Knierem, Hayes & Ware (2006)
- Random sample of tests/items from programs of all types throughout the US
Thank you!
References


References


References


References


References


References


Dambrot, F. (1980). Test item order and academic ability, or should you shuffle the test item deck? *Teaching of Psychology, 7*(2), 94-96.
References


References


References


References


References


References


References


References


doi:10.1097/NNE.0b013e3181aaba94

References


References


References


References


References
