MEASURING INTERVENTION EFFECTIVENESS: THE BENEFITS OF AN ITEM RESPONSE THEORY APPROACH

Society for Research in Educational Effectiveness
September 6, 2012

Katherine McEldoon, Sun-Joo Cho, & Bethany Rittle-Johnson

Department of Psychology & Human Development
Peabody College, Vanderbilt University
Intervention Research Design

Roadmap

• Issues with RM ANOVA
• Item Response Theory (IRT) Models
  • Overcome limitations of RM ANOVA
  • Offer additional benefits
• GEL MIRT Model as alternative to RM ANOVA
Control Posttest Retention

Treatment Posttest Retention

Between Conditions

Within Repeated Measures

• Between-Within (split-plot) Design
Intervention Research Design

- Binary Scores
- Proportion Correct
- Assessment Subcomponents
- Multidimensionality

1. \( 8 + 4 = [\_] + 5 \)  
   - 6
   - ✔️ 1
   - Procedural Knowledge

2. \( 7 + 6 + 4 = 7 + [\_] \)  
   - 24
   - ✗ 0

3. \( 3 + 5 = 5 + 3 \)  
   - True or False

4. What does the equal sign mean?  
   - The answer goes next.
Data Structure for Analysis

- Posttest
- Retention
- Posttest
- Retention
Data Structure for Analysis

- Posttest
  - Control
    - Item Responses
  - Treatment
    - Item Responses
  - Proportion Correct
    - Conceptual Proportion Correct
  - Proportion Correct
    - Conceptual Proportion Correct

- Retention
  - Control
    - Item Responses
  - Treatment
    - Item Responses
  - Proportion Correct
    - Conceptual Proportion Correct
  - Proportion Correct
    - Conceptual Proportion Correct
Data Structure for Analysis

<table>
<thead>
<tr>
<th>Condition</th>
<th>Subject</th>
<th>Conceptual Item Responses</th>
<th>Procedural Item Responses</th>
<th>Concept. P. Correct</th>
<th>Proced. P. Correct</th>
<th>Total Proportion Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1</td>
<td>1 0 0</td>
<td>0 1 0</td>
<td>.33</td>
<td>.33</td>
<td>.33</td>
</tr>
<tr>
<td>Control</td>
<td>2</td>
<td>1 1 0</td>
<td>1 1 1</td>
<td>.66</td>
<td>1.0</td>
<td>.83</td>
</tr>
<tr>
<td>Control</td>
<td>3</td>
<td>0 1 0</td>
<td>1 1 0</td>
<td>.33</td>
<td>.66</td>
<td>.50</td>
</tr>
<tr>
<td>Control</td>
<td>4</td>
<td>1 1 0</td>
<td>1 0 0</td>
<td>.66</td>
<td>.33</td>
<td>.50</td>
</tr>
<tr>
<td>Treatment</td>
<td>5</td>
<td>1 1 0</td>
<td>1 0 1</td>
<td>.66</td>
<td>.66</td>
<td>.66</td>
</tr>
<tr>
<td>Treatment</td>
<td>6</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Treatment</td>
<td>7</td>
<td>1 0 0</td>
<td>1 1 0</td>
<td>.33</td>
<td>.66</td>
<td>.50</td>
</tr>
<tr>
<td>Treatment</td>
<td>8</td>
<td>0 1 1</td>
<td>1 1 0</td>
<td>.66</td>
<td>.66</td>
<td>.66</td>
</tr>
</tbody>
</table>

Binary Responses to Items

ANOVA Dependent Variables
Issues with RM ANOVA Analysis

1. Samples used in intervention research often violate RM ANOVA assumptions

2. Proportion correct scale not an interval scale, and does not allow for meaningful comparisons
Issues with RM ANOVA - Assumptions

1. Samples used in intervention research often violate RM ANOVA assumptions

RM ANOVA Assumptions:
   A. Independence
   B. Normality
   C. Equality of Variances

Although relatively robust, when violated, conclusions can be biased (Dixon, 2008; Embretson, 1991; Jaeger, 2008)
Issues with RM ANOVA - Assumptions

A. Independence
   - Violated by hierarchical and nested structure of educational settings
Issues with RM ANOVA - Assumptions

B. Normality

• Between: Conditions
• Within: Timepoints

• Distributions often not normal
  • Bimodal distributions due to differential intervention effects

• When group sizes are unequal, results biased (Wilcox, 2005)
Issues with RM ANOVA - Assumptions

C. Equal Group Variances Between
   • Between: Conditions
   • Within: Timepoints

   • Small violations can lead to inflated Type I error rates (Boik, 1981)

   \[
   \text{Var} = .022 \\
   \text{Var} = .083 \\
   \text{Var} = .084
   \]
Issues with RM ANOVA – Interval Scale

Proportion correct is not empirically an interval scale
Meaningful comparisons can only be made from the same initial value
Can lead to biased results (Agresti, 2002; Dixon, 2008)
An Alternative: IRT Models

• Item Response Theory Models
  
  • Overcomes limitations of RM ANOVA models
    • No assumptions of independence, normality, or equal variance
      • However, does have assumptions of dimensionality and local independence
    • Interval scale, allowing for meaningful comparisons
  
  • Has additional benefits
    • Incorporates more informative metrics
    • Accounts for measurement error
Advances in IRT Models

• IRT Models in the past have been used for:
  • Individual Differences
  • Large-scales tests

• Recent Advances:
  • Individual & Group Differences
  • Smaller sample sizes typical of education research
    • E.g. 100 subjects and 20 assessment items
    • New estimation methods
      (Random item approach, Cho & Rabe-Hesketh, 2011, 2012)
  • Simultaneously Handle:
    • Longitudinal Designs
    • Multidimensional Constructs
# Data Structure for Analysis

<table>
<thead>
<tr>
<th>Condition</th>
<th>Subject</th>
<th>Conceptual Item Responses</th>
<th>Procedural Item Responses</th>
<th>Concept. P. Correct</th>
<th>Proced. P. Correct</th>
<th>Total Proportion Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1</td>
<td>1 0 0</td>
<td>0 1 0</td>
<td>.33</td>
<td>.33</td>
<td>.33</td>
</tr>
<tr>
<td>Control</td>
<td>2</td>
<td>1 1 0</td>
<td>1 1 1</td>
<td>.66</td>
<td>1.0</td>
<td>.83</td>
</tr>
<tr>
<td>Control</td>
<td>3</td>
<td>0 1 0</td>
<td>1 1 0</td>
<td>.33</td>
<td>.66</td>
<td>.50</td>
</tr>
<tr>
<td>Control</td>
<td>4</td>
<td>1 1 0</td>
<td>1 0 0</td>
<td>.66</td>
<td>.33</td>
<td>.50</td>
</tr>
<tr>
<td>Treatment</td>
<td>5</td>
<td>1 1 0</td>
<td>1 0 1</td>
<td>.66</td>
<td>.66</td>
<td>.66</td>
</tr>
<tr>
<td>Treatment</td>
<td>6</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Treatment</td>
<td>7</td>
<td>1 0 0</td>
<td>1 1 0</td>
<td>.33</td>
<td>.66</td>
<td>.50</td>
</tr>
<tr>
<td>Treatment</td>
<td>8</td>
<td>0 1 1</td>
<td>1 1 0</td>
<td>.66</td>
<td>.66</td>
<td>.66</td>
</tr>
</tbody>
</table>

**IRT Dependent Variables**

**ANOVA Dependent Variables**
IRT Model

Student Ability

Item Difficulty

High

Low

Mean: 0

Student Ability: 1.1
Item Difficulty: 1.1
Probability = .50

Item Difficulty: 2.3
Probability = .24

Item Difficulty: -1.3
Probability = .91
Additional Benefits of IRT Models

1. Less sensitive to violations of RM ANOVA assumptions
2. Interval scale allows for meaningful comparisons
3. More informative metrics of student ability and item difficulty
4. Latent variable separates true group difference from measurement error
### Additional Benefits of IRT Models

3. More informative metrics of student ability and item difficulty
   - Ability estimates for each subscale
   - Correlation structure between subscales

<table>
<thead>
<tr>
<th>Item Difficulty</th>
<th>Student A</th>
<th>Student B</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.4</td>
<td>✔️ 1</td>
<td>✔️ 1</td>
</tr>
<tr>
<td>.96</td>
<td>✔️ 1</td>
<td>1.2</td>
</tr>
<tr>
<td>-.33</td>
<td>✗️ 0</td>
<td>✗️ 0</td>
</tr>
<tr>
<td>.45</td>
<td>✗️ 0</td>
<td>✔️ 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Procedural</th>
<th>Conceptual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>2.</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>3.</td>
<td>✗️</td>
<td>✗️</td>
</tr>
<tr>
<td>4.</td>
<td>✗️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

1. $8 + 4 = \square + 5$
2. $7 + 6 + 4 = 7 + \square$
3. $3 + 5 = 5 + 3$
4. True or False
5. What does the equal sign mean?
Additional Benefits of IRT Models

4. Latent variable separates true group difference from measurement error

Latent Variable: not directly observed, but inferred from other variables that are observed

IRT: Latent Ability = Latent Ability Estimate + Standard Error

ANOVA: Observed Score = True Score + Error
### Data Structure for Analysis

<table>
<thead>
<tr>
<th>Condition</th>
<th>Subject</th>
<th>Conceptual Item Responses</th>
<th>Procedural Item Responses</th>
<th>Concept. P. Correct</th>
<th>Proced. P. Correct</th>
<th>Total Proportion Correct</th>
<th>Concept Ability Est (se)</th>
<th>Proced. Ability Est (se)</th>
<th>Total Ability Est (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1</td>
<td>1 0 0</td>
<td>0 1 0</td>
<td>.33</td>
<td>.33</td>
<td>.33</td>
<td>-.42 (.12)</td>
<td>-1.3 (.11)</td>
<td>-.86 (.14)</td>
</tr>
<tr>
<td>Control</td>
<td>2</td>
<td>1 1 0</td>
<td>1 1 1</td>
<td>.66</td>
<td>1.0</td>
<td>.83</td>
<td>.15 (.13)</td>
<td>1.1 (.08)</td>
<td>.63 (.11)</td>
</tr>
<tr>
<td>Control</td>
<td>3</td>
<td>0 1 0</td>
<td>1 1 0</td>
<td>.33</td>
<td>.66</td>
<td>.50</td>
<td>-.42 (.09)</td>
<td>.58 (.10)</td>
<td>.08 (.09)</td>
</tr>
<tr>
<td>Control</td>
<td>4</td>
<td>1 1 0</td>
<td>1 0 0</td>
<td>.66</td>
<td>.33</td>
<td>.50</td>
<td>.15 (.08)</td>
<td>-1.3 (.13)</td>
<td>-.58 (.12)</td>
</tr>
<tr>
<td>Treatment</td>
<td>5</td>
<td>1 1 0</td>
<td>1 0 1</td>
<td>.66</td>
<td>.66</td>
<td>.66</td>
<td>.15 (.08)</td>
<td>.58 (.06)</td>
<td>.37 (.11)</td>
</tr>
<tr>
<td>Treatment</td>
<td>6</td>
<td>1 1 1</td>
<td>1 1 1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.3 (.11)</td>
<td>1.1 (.08)</td>
<td>1.2 (.18)</td>
</tr>
<tr>
<td>Treatment</td>
<td>7</td>
<td>1 0 0</td>
<td>1 1 0</td>
<td>.33</td>
<td>.66</td>
<td>.50</td>
<td>-.42 (.13)</td>
<td>.58 (.07)</td>
<td>.08 (.07)</td>
</tr>
<tr>
<td>Treatment</td>
<td>8</td>
<td>0 1 1</td>
<td>1 1 0</td>
<td>.66</td>
<td>.66</td>
<td>.66</td>
<td>.15 (.07)</td>
<td>.58 (.07)</td>
<td>.37 (.10)</td>
</tr>
</tbody>
</table>
IRT Model
GEL MIRT Alternative to RM ANOVA

Cho, Athay, & Preacher (2012)
GEL MIRT Alternative to RM ANOVA

Student Ability

Treatment

Difference of Abilities Between Conditions

Difference of Difficulties Between Item Groups

Conceptual

Procedural

Control

Cho, Athay, & Preacher (2012)
The Dataset

156 2nd & 3rd Grade Students

Control
N = 79

Conceptual Instruction
Problem Solving with accuracy feedback

Problem Solving: Math Equivalence

2 + 5 + 8 = ____ + 8
3 + 4 + 6 = ____ + 4

Treatment
N = 77

Problem Solving with accuracy feedback
Conceptual Instruction

Immediate Posttest
2 Week Retention

Matthews, Rittle-Johnson, McEldoon & Taylor, 2012; Rittle-Johnson, Matthews, Taylor & McEldoon, 2011

Conceptual
3 + 5 = 5 + 3  T/F
What does the equal sign mean?

Procedural
8 + 4 = [□] + 5
7 + 6 + 4 = 7 + [□]

~ 20 minutes
Comparing Output – RM ANOVA and GEL MIRT

<table>
<thead>
<tr>
<th>Source</th>
<th>RM ANOVA</th>
<th>GEL MIRT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.(p)</td>
</tr>
<tr>
<td>Intercept</td>
<td>1298.2</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Condition</td>
<td>0.592</td>
<td>0.443</td>
</tr>
<tr>
<td>Time</td>
<td>15.02</td>
<td>&lt;.01*</td>
</tr>
<tr>
<td>Time X Condition</td>
<td>0.54</td>
<td>0.464</td>
</tr>
<tr>
<td>Error (SS)</td>
<td>16.2</td>
<td></td>
</tr>
</tbody>
</table>
Comparing Output – RM ANOVA and GEL MIRT

Item Group Differences

<table>
<thead>
<tr>
<th>Source</th>
<th>RM ANOVA</th>
<th>GEL MIRT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.(p)</td>
</tr>
<tr>
<td>Intercept</td>
<td>1298.2</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Condition</td>
<td>0.592</td>
<td>0.443</td>
</tr>
<tr>
<td>Time</td>
<td>15.02</td>
<td>&lt;.01*</td>
</tr>
<tr>
<td>Time X Condition</td>
<td>0.54</td>
<td>0.464</td>
</tr>
<tr>
<td>Item Group</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>Error</td>
<td>(SS) 11.1</td>
<td></td>
</tr>
</tbody>
</table>
Comparing Output – RM ANOVA and GEL MIRT Conceptual Subscale

<table>
<thead>
<tr>
<th>Source</th>
<th>RM ANOVA</th>
<th>GEL MIRT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.(p)</td>
</tr>
<tr>
<td>Intercept</td>
<td>1231.9 &lt;.01</td>
<td>0.488</td>
</tr>
<tr>
<td>Condition</td>
<td>3.169 0.077</td>
<td>0.32</td>
</tr>
<tr>
<td>Time</td>
<td>1.134 0.289</td>
<td>0.188</td>
</tr>
<tr>
<td>Error</td>
<td>(SS) 3.123</td>
<td>Individual Student Ability Estimates</td>
</tr>
</tbody>
</table>
Support from Simulation Studies

• Simulation studies support that IRT models are more accurate at detecting true group differences than RM ANOVA Models

• When:
  • True group differences on latent variable
  • RM ANOVA assumptions are violated

• Detection Rates of Group Differences:
  • RM ANOVA: 44%
  • GEL MIRT: 99%

Conclusions

• Researchers should consider the advantages of an IRT approach for evaluating intervention effectiveness
  • GEL MIRT model (Cho, Athay, & Preacher, 2012)

• Pro:
  • More informative metrics
  • Less prone to biased results
  • Can be performed using the open-source and free program R
    • Details of the model, as well as information how to run these analyses can be found in Cho, Athay, & Preacher (2012)

• Con:
  • Requires more technical proficiency on the part of the data analyst
  • Challenging to understand
Thank You

Sun-Joo Cho & Bethany Rittle-Johnson
Michael Nelson & Marci DeCaro
Children’s Learning Lab

Vanderbilt Children’s Learning Lab
http://peabody.vanderbilt.edu/departments/psych/research/research_labs/childrens_learning_lab/index.php

GEL MIRT Model Paper and Details
http://quantpsy.org/pubs.htm

The first author is supported by a predoctoral training grant provided by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040110 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.