Abstract Title Page

Title:

Measuring Intervention Effectiveness: The Benefits of an Item Response Theory Approach

Authors and Affiliations:

Katherine McEldoon, Peabody College, Vanderbilt University

Sun-Joo Cho, Peabody College, Vanderbilt University

Bethany Rittle-Johnson, Peabody College, Vanderbilt University

Abstract Body

Background / Context:

Assessing the effectiveness of educational interventions relies on quantifying differences between interventions groups over time in a between-within design[1]. Binary outcome variables (e.g., correct responses versus incorrect responses) are often assessed. Widespread approaches use percent correct on assessments, and repeated measures analysis of variance (ANOVA) methods to detect differences between groups. However, this approach is not ideal, as in fact several assumptions are often violated when using this method that can result in less informative and at times biased and spurious findings (Dixon, 2008; Embretson, 1994). An alternative approach is to utilize item response models to detect differences between intervention groups over time.

Purpose / Objective / Research Question / Focus of Study:

The quantification of change in ability due to learning, in both individuals and between groups, is the primary outcome measure of educational intervention research. As such, it is important that these measures are accurate and informative. The specific benefits of one-parameter item response models for intervention research are presented and contrasted with repeated measures ANOVA. The appropriateness of these models will be demonstrated using data from elementary students who participated in a tutoring intervention on mathematical equivalence. Advances in the development of item response models now make them a viable option for accommodating educational datasets, in terms of practical constrains of sample sizes and the nature of the samples. Item response models can accommodate longitudinal designs with repeated measures (Embretson, 1991), multidimensional constructs (Briggs & Wilson, 2003), and a new generalized explanatory longitudinal item response model for multidimensional tests (Cho, Athay, & Preacher, in press) can accommodate both, as well as quantify and significance test the effect of intervention condition much like ANOVA approaches.

The benefits of item response methodology for intervention research will be contrasted with repeated measures ANOVA approaches, using a longitudinal intervention dataset having a between-within design from elementary students learning about mathematical equivalence. The dependent measures of percent correct in repeated measures ANOVA approaches and item responses in item response models will be contrasted, as well as the methods for quantifying differences between groups using repeated measures ANOVA approaches and item response models.

Intervention / Program / Practice:

Second and third grade students who scored below 75% correct on the pretest participated in a 20-minute one-on-one tutoring intervention that focused on mathematical equivalence problems. Students were from urban and suburban schools in a Southeastern city. Students then completed

[1] A between-within design is also called a split-plot design, which is from in agricultural research, or a mixed design. The term “mixed” is used because the design mixes between-subjects and within-subjects factors, not because it has both random and fixed effects in linear or nonlinear mixed models.

an immediate posttest, and a two-week retention test that took about 35 minutes to complete. Span from pretest to retention test averaged about 4 weeks. The assessments have been developed and tested for reliability and validity in prior work (Matthews, Rittle-Johnson, McEldoon & Taylor, 2012; Rittle-Johnson, Matthews, Taylor & McEldoon, 2011). The assessment has two components or dimensions that assess students' procedural and conceptual knowledge. Models will be demonstrated from studies utilizing this method with samples of 347 and 151 students.

Statistical, Measurement, or Econometric Model:

The models described here are from the one-parameter item response theory (IRT) family (Birnbaum, 1968). A one-parameter item response model considers both respondent ability and item difficulty simultaneously on the same interval scale, modeling the probability that a particular respondent will answer a particular item correctly (Birnbaum, 1968). Therefore, a student's ability parameter takes into account the difficulty of the specific items they correctly answered. This ability parameter is an important feature that raw score approaches such as in a repeated measures ANOVA models do not utilize. The benefits of item response models for measuring individual change will be demonstrated using longitudinal intervention data from 347 students fitted to the Embretson's multidimensional latent trait model for measuring learning and change (1991), and will be compared to percent correct scores. Additionally, a new generalized explanatory longitudinal item response model for multidimensional tests (Cho, Athay, & Preacher, in press) will be described and contrasted with repeated measures ANOVA methods using a sample with 151 students. This new model can quantify the students' change in ability on specific assessment subscales and quantify the effect of intervention condition, much like an ANOVA model.

Usefulness / Applicability of Method:

Problems with ANOVA Approaches and an Item Response Model as its Alternative

1. *IRT ability estimates are more accurate and informative than total scores or percent correct scores, which are typical of traditional ANOVA approaches.*

A normal model such as ANOVA approaches assumes that the outcome variables are measured at least on the interval scale and are continuous. Linear models such as ANOVA approaches for categorical response data have been known problematic for a long time (for summaries, see Agresti, 2002, p. 120). The proportion correct is constrained to the range 0-1. This constrained range can lead problems in using ANOVA as described in Agresti (2002) and Dixon (2008). Item response models utilize an interval scale, where the distance between all score points are constant. Classical test models such as ANOVA can have interval scale properties, but only when the data is normally distributed, and this is often not the case with intervention data. This results in the relative distance between scores not being constant in classical test scaling. This is a crucial point for intervention research, as comparisons and change scores can only be meaningful when there is an interval scale. If there is not, comparisons can only properly be made when the initial score values are the same. Because of the compression of the total scores or percent correct scales, small changes from a high score means more than a small change from

a moderate score. For example, a student who made a 19% pre- to post-test gain from 10% to 29% had a 0.85 gain in ability estimate, whereas a student who also made a 19% gain from 62% to 81% had a much larger gain of 1.57 in their ability estimate. In item response models, the relative distance between scores is always constant and can be meaningfully interpreted no matter what the time point, ability estimate score, or intervention condition. As such, item response models are ideally suited for the accurate measurement and comparison of learning over time and between groups.

2. *Item response model takes the measurement error into account appropriately by modeling a latent variable so that the model allows investigators to separate true group difference from measurement error.*

Item response models make latent ability estimates, instead of simply using total score or percent correct. Since they are an estimate, the have standard error scores, and can also be used to make predictions about student performance on future items. ANOVA does not address the issue of measurement error in outcome variables. It uses an observable outcome variable, which ignores the measurement error.

3. *Item response models produce a rich ability estimate for each student for each subscale.*

Researchers may want to consider student abilities on different assessment subscales, because students may have differential performance across different dimensions, or item types. Importantly, knowledge of these different dimensions might be differentially affected by instruction intervention, and so it would be important to detect and quantify these changes in knowledge after intervention. Multidimensional item response models can accommodate this (Briggs & Wilson, 2003; Cho et al., in press). When using a multidimensional model, we still retain information about the students' performance on the other subscales (and their correlational structure) when estimating a student's ability on a particular subscale. Although multivariate ANOVA (MANOVA) allows analyzing multiple outcome variables simultaneously, it uses a single outcome (e.g., proportion correct for binary response) so items that actually measure different domains or facets of constructs are aggregated and are erroneously treated as unidimensional.

Item Response Models More Accurately Quantify Group Differences Over Time: Repeated Measures ANOVA vs. Generalized Explanatory Longitudinal Item Response Models

When looking to quantify differences between experimental groups in a longitudinal design, repeated measures ANOVA methods are used often. Despite their prevalence, repeated measures ANOVA models are not ideal for use with intervention data. Due to the nature of intervention research, repeated measures ANOVA assumptions are often not met, and this can lead to biased conclusions **(**Dixon, 2008; Embretson, 1991; Jaeger, 2008). The first assumption that must be met is that the samples are **independent**. Educational settings have systematic hierarchical structure and natural nesting (e.g. school district, school, grade, classroom), clearly violating this condition. **Normality** is also assumed. However, distributions are often not normal, particularly at posttest, where distributions may be bimodal due to the intervention being effective for some students but not for others. In the example dataset containing 347 students, the distribution of the

pretest data is skewed, and the post and retention test distributions clearly have a bimodal shape (Figures 1-3). The Shapiro-Wilks test for normality indicates that all three distributions are significantly different from normal (Pre, $t(347) = 0.945$, $p < .001$; Post $t(347) = 0.927$, $p < .001$; Reten $t(347) = 0.914$, $p < .001$). Because the distributions of the post and retention tests are bimodal, a transformation cannot easily be used to correct this. In addition to being a violation of the repeated measures ANOVA approach, this non-normality also results in biased change score interpretations, as discussed earlier. The last assumption is of **homoscedasticity and sphericity**. This assumption is frequently violated in intervention data due to the variance often being greater after intervention because of differential effects of intervention on individuals. Indeed, this is the case in the example data as well. Levene's test for homogeneity of variance is not met between pretest and post ($t(1,692) = 250.58$, $p < .001$) or retention test ($t(1,692) = 251.04$, $p < .001$), however it is equal between post and retention ($t(1,692) = 0.022$, $p = .881$). Even small violations of homoscedasticity and sphericity can lead to greatly inflated Type I error rates for both omnibus tests and contrasts (Boik, 1981). Because intervention data will typically violate many of these assumptions, as the current data set has, the repeated measures ANOVA approaches are not ideal. Indeed, inappropriately used the repeated measures ANOVA approach can lead to biased results and spurious interactions (Dixon, 2008; Embretson, 1991; Jaeger, 2008).

Until recently, there were no direct alternatives that would answer the same research questions as ANOVAs within item response models; such as quantifying the effect of intervention, time, and providing significance tests for these factors. Item response models require that items are locally independent and that each subscale being measured is unidimensional. These requirements are easily tested and accommodated within intervention research designs. A new generalized explanatory longitudinal item response model for multidimensional tests (Cho et al., in press) provides such an alternative. Suitable for sample sizes as small as 100 students and 20 assessment items, this model provides a quantification of effect of intervention condition, of time, and time by intervention condition interactions, as well as a significance tests for each. These results are much like those provided in a repeated measures ANOVA, however, all of the benefits of more informative and accurate ability parameters are built into this model and problematic issues with ANOVA models are avoided. See the results of this generalized explanatory model (Model 1) run with our sample of 151 students in Appendix C. Here we can look for an effect of condition in the fixed effects section of the results to see that our intervention condition was 0.129 logits lower on their ability parameter than the control condition, but this difference was not significant ($p = 0.406$).

Often researchers want to investigate the effect of intervention condition on various intervention subscales. In the current dataset, the assessment of mathematical equivalence is broken down into a procedural and conceptual knowledge section. The model developed by Cho et al. also estimates the difficulty, or effect of the subscales, which could be of value to intervention researchers. In Model 2, the fixed effect of 'CONvPROC' quantifies the relative difficulty of the conceptual knowledge items relative to the procedural knowledge items. We can interpret that the conceptual knowledge item difficulty is 1.11 logits smaller (or easier) than the procedural knowledge items (however $p = 0.123$). Additionally, researchers may be interested in determining if there is an item subscale by condition interaction, to see if the intervention condition performed differently than the control on different subscales. The subscale by condition interaction (Con.v.Proc*Condition) indicates that the conceptual knowledge items

were 0.352 logits easier for the intervention condition than the control, but this difference was not significant ($p = 0.310$). These generalized explanatory item response models offer ways of quantifying and significance testing various differences between conditions and item subscales that repeated measures ANOVA models cannot.

This new generalized explanatory longitudinal item response model for multidimensional tests provides intervention researchers a more informative, less biased, and equally powerful method for evaluating intervention effectiveness.

Conclusions:

In conclusion, item response models offer many methodological advantages in the quantification of individual learning and group change over time compared to repeated measure ANOVA approaches based on percent correct outcomes. In particular, the generalized explanatory longitudinal item response model for multidimensional tests (Cho et al., in press) quantifies and tests for differences between intervention conditions, while utilizing the more informative and less problematic metrics of student performance. In addition to being methodologically more sound, these analyses 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 et al. (in press). One drawback to preforming IRT analyses is that they do require more technical proficiency on the part of the data analyst than ANOVA approaches. Nevertheless, researchers should strive to adapt this more informative and less biased metric in the evaluation of intervention effectiveness.

Appendices

Appendix A. References

Agresti, A. (2002). *Categorical Data Analysis (Wiley Series in Probability and Statistics)* (Second Edition, p. 710). Hoboken, New Jersey: John Wiley and Sons.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Boik, R. J. (1981). Priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, *46*(3), 241-255.

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). Mahwah, NJ: Erlbaum.

Briggs, D. C., & Wilson, M. (2003). An Introduction to Multidimensional Measurement using Rasch Models. *Journal of Applied Measurement*, *4*(1), 87-100.

Cho, S.-J., Athay, M., & Preacher, K. J. (in press). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology*.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447-456. doi:10.1016/j.jml.2007.11.004

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. Psychometrika, 56, 495-515. doi:10.1007/BF02294487

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434-446. doi:10.1016/j.jml.2007.11.007

Matthews, P., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education*, *43*(3).

Rasch, G. (1980). Probabilistic model for some intelligence and attainment tests. Chicago, IL: University of Chicago Press.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology*, *103*(1), 85-104. doi:10.1037/a0021334

Appendix B. Figures

Figure 1. Pretest Score Distribution

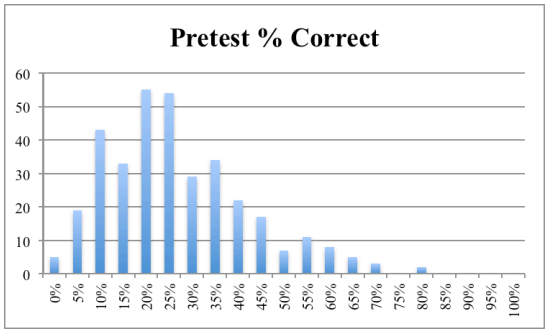


Figure 2. Posttest Score Distribution

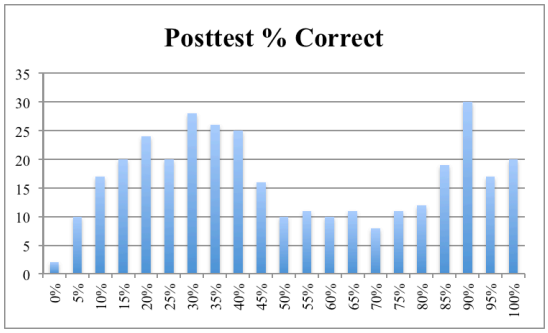
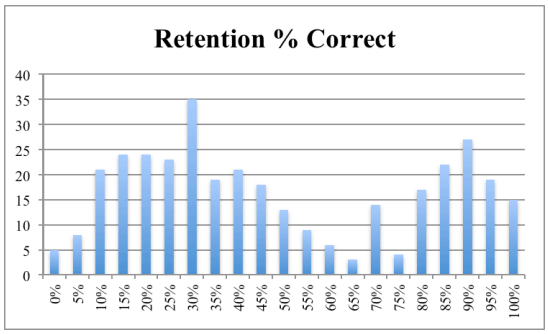


Figure 3. Retention Test Score Distribution



Appendix C. Generalized Explanatory IRT Model R Code and Select Results

```
# Load Data to R
data <- read.table("C:/Path/data.txt", header=T, fill=T)

#Call libraries
library(lme4)

# Establishing Matrices
data$ITEM <- as.factor(data$Item)
data$TIME <- as.factor(data$Time)
data$CONDITION <- as.factor(data$Condition)
data$CONvPROC <- as.factor(data$Con)
data$CONT1 <- (data$Con)*(data$Q1)
data$CONT2 <- (data$Con)*(data$Q2)
data$CONT3 <- (data$Con)*(data$Q3)
data$PROCT1 <- (data$Proc)*(data$Q1)
data$PROCT2 <- (data$Proc)*(data$Q2)
data$PROCT3 <- (data$Proc)*(data$Q3)

# Code for Model 1: Generalized Explanatory IRT Model that Parallels Repeated Measures ANOVA
GenExplanIRTModel <- lmer(Resp ~ 1 + TIME*CONDITION + CONvPROC + (1|Item) + (ConceptualT1 +
ConceptualT 2+ ConceptualT 3-1|Person)+(ProceduralT1+ ProceduralT2+ ProceduralT3 -1|Person), data,
binomial("logit"))
GenExplanIRTModel

#Results for Model 1
Random effects:
 Groups Name        Variance  Std.Dev. Corr
 Person  PROCT1      3.032   1.741
         PROCT2      3.948   1.986    -0.050
         PROCT3      1.157   1.076    -0.183   -0.191
 Person  CONT1       0.691   0.831
         CONT2       0.558   0.747     0.407
         CONT3       0.033   0.183    -0.968   -0.163
 item1   (Intercept) 2.493   1.579
Number of obs: 9300, groups: Person, 155; item1, 20

Fixed effects:
                    Estimate   Std. Error z value   Pr(>|z|)
(Intercept)         -0.2974    0.3728     -0.798    0.4250
TIME2               1.4109     0.1335     10.569    <2e-16 ***
TIME3               1.5883     0.1344     11.815    <2e-16 ***
CONDITION           -0.1296    0.1559     -0.831    0.4057
TIME2:CONDITION     0.2644     0.1872     1.412     0.1579
TIME3:CONDITION     0.4245     0.1893     2.243     0.0249 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
             (Intr)   TIME2   TIME3   CONDIT TIME2:
TIME2        -0.060
TIME3        -0.086   0.658
CONDITION    -0.210   0.144   0.208
TIME2:CONDI  0.043    -0.704  -0.460  -0.206
```

TIME3:CONDI 0.061 -0.457 -0.698 -0.295 0.654

 # Model 2: Code for Generalized Explanatory IRT Model with Item Type (Conceptual) by Condition Interaction

```
GenExplanIRTModelwItemTypeInteraction <- lmer(Resp ~ 1 + TIME + CONvPROC +
CONvPROC*CONDITION + (1|Item) + (ConceptualT1 + ConceptualT2+ ConceptualT3-
1|Person)+(ProceduralT1+ ProceduralT2+ ProceduralT3 -1|Person), data, binomial("logit"))
GenExplanIRTModelwItemTypeInteraction
```

#Select Results for Model 2

Fixed effects:

	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	0.26568	0.52602	0.505	0.614
Time2	1.54647	0.09523	16.240	<2e-16 ***
Time3	1.80560	0.09710	18.595	<2e-16 ***
Conceptual(vs Procedural)	**-1.11337**	**0.72175**	**-1.543**	**0.123**
Condition	-0.29204	0.30188	-0.967	0.333
Con.v.Proc*Condition	**0.35247**	**0.34697**	**1.016**	**0.310**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1