Measurement Error in Hierarchical Gain Score Model

Changhui Zhang and Raven McCrory
Michigan State University

This paper compares three approaches for solving the problem of measurement error in a hierarchical gain score model. The pre-test and post-test scores are IRT scores with measurement error. Explanatory variables at student level and class level are considered in the model. Simulation results show that the gain score model that does not consider measurement error overestimates the explanatory variable effects, while the growth model using IRT estimation error underestimates the effects. The Markov Chain Monte Carlo method generates the best estimation.

Keyword: Measurement error, Bayesian, Item Response Theory, Hierarchical model

Purpose of the Study

The purpose of this study is to compare HLM2, HLM3 Growth Model, and MCMC procedures in dealing with measurement error in hierarchical gain score models. Two criteria are used to compare the estimators. One is whether they are biased. An estimator is unbiased if the expected value of the estimator equals the true parameter. The other is efficiency which is measured by the standard error of the estimator.

Theoretical Framework

A gain score, which is defined as the difference between post-test score and pre-test score, is often used to measure progress in education studies. Often, the researchers are interested in finding the factors that influence this student gain. However, test scores are not precise measures and researchers using gain scores often ignore the associated measurement error. It is not clear how this neglect affects the estimation of the effects of predictive factors. This paper explores the measurement error effect in hierarchical linear models (HLM) using IRT scores and then compares three different approaches to solving the measurement error problem.

Measurement error is the difference between the observed score and the true score. In an Item Response Theory (IRT) framework, the student ability, T_i, is assumed to be on a latent continuum. The IRT score, X_i, the maximum likelihood estimator of T_i, obtained based on the answer patterns, is an observed score. Thus

$$X_i = T_i + E_i$$

Where E_i is the measurement error.

Xi is a consistent estimator and asymptotically μ_{E_i} approaches 0 as the sample size increases. The variance of the error is calculated as follows:

$$Var(E_i) = Var(X_i \mid T_i) = \frac{1}{I(T_i)}$$

where $I(T_i)$ is the test information function for X_i.

A more detailed explanation of information function can be found in Lord (1980, p. 71). The important implications here are that the error variance is conditional on T_i and the errors are heterogeneous.

Data sources

This work is part of a larger study that explores the mathematics content taught to undergraduate prospective elementary teachers (PSTs) at 40 universities in four states. PSTs took pre and post-tests to assess their mathematical knowledge. Instructors of these classes were also given a survey. For more information about the project, including student test data and instructor surveys, see [authors, 2008]. The student and instructor data were then entered into a model with students nested in classes with a specific teacher.

Method

After looking at the models, three specific variables seemed to explain variance of gain scores. The explanatory variable at the student level was ACT score and the explanatory variables at the teacher level were teaching method and textbook. Teaching method is related to how much a teacher claimed that they lectured versus how much the students was expected to contribute to the discussion in class. ACT scores are an indicator of general achievement before entering college. The variable related to the textbook was entered as a dichotomous variable where 1 indicated that the teacher used one the primary textbooks (designed specifically for a mathematics class for teachers) and 0 indicated some other textbook or no textbook at all.

There are several ways to estimate the effects of these explanatory variables. Three such approaches are described below.

HLM2 Procedure

If the measurement error is ignored, a 2-level HLM model can be used to estimate the covariate coefficients.

Level-1 Model
Gain = B0 + B1*(ACT) + R
Level-2 Model
B0 = G00 + G01*(METHOD) + G02*(TEXT) + U0
B1 = G10

Where R is the variation of gain score within each class.

HLM3 Growth Model Procedure

One possible solution to the question is to setup a 3-level HLM growth model as follows:

Level-1 Model
    Test_score = P0 + P1*(TIME) + E
Level-2 Model
    P0 = B00 + B01*(ACT) + R0
    P1 = B10 + B11*(ACT) + R1
Level-3 Model
    B00 = G000 + U00
    B01 = G010
    B10 = G100 + G101(METHOD) + G102(TEXT) + U10
    B11 = G110

Where in level 1, E is the measurement error, fixed to equal the estimation error in IRT estimation procedures. Thus it is handled as a variance known problem at level 1. More explanation of variance in HLM can be found in Raudenbush and Bryk (2002).

MCMC Procedure

In the previous two procedures, the test scores were estimated by using IRT models. The measurement error was obtained from IRT estimation procedures. It is possible to combine IRT score estimation with hierarchical models. So the model can be written as:

Level-1 IRT Model
    Prob(Y=1) = P
    log[P/(1-P)] = A*(P0 + P1*(TIME)- B)
Level-1 Structural Model
    P0 = B00 + B01*(ACT) + R0
    P1 = B10 + B11*(ACT) + R1
Level-2 Model
    B00 = G000 + U00
    B01 = G010
    B10 = G100 + G101(METHOD) + G102(TEXT) + U10
    B11 = G110

Where A and B are item parameters and Y is the 0 or 1 response to a particular item. Now the model cannot be solved by using HLM software. Markov Chain Monte Carlo (MCMC) methods are used to estimate the parameters in the model. MCMC methods produce simulated chain values in which each of the values are mildly dependent on the preceding one. Once the chain has run sufficiently long enough, the summary statistics of the simulated values can be used to estimate the parameters. Interested readers can find more discussions on MCMC in Gill (2004). Particularly, the rational of using Bayesian method to solve IRT and HLM combination models can be found in Fox and Glas (2001).

In this study, the MCMC chains are produced by using WinBUGS software (Medical Research Council Biostatistics Unit, Cambridge, www.mrc-bsu.cam.ac.uk/bugs).

To test which method most accurately predicted the results, forty classes with thirty students in each class were generated. The generated data included a pre-test of 26 items and a post-test of 20 items for each student, as in the actual project data. The item parameters were taken from the Learning for Mathematics for Teaching Project (LMT) which developed and piloted all the items, generating IRT parameters in a two-parameter model. The parameters were then assumed to be known for data generation and later in student ability estimation. The student responses were generated based on a two parameter logistic regression IRT model. True student abilities were generated by using a two-level linear model as in the following

Level-1 IRT Model

$\mathrm{Prob}(Y=1) = P$

$\log[P/(1-P)] = A*(P0 + P1*(TIME)- B)$

Level-1 Structural Model

$P0 = B00 + B01*(ACT) + R0$

$P1 = B10 + B11*(ACT) + R1$

Level-2 Model

$B00 = 50.00 + U00$

$B01 = 1.00$

$B10 = 4.30 + 3.00*(METHOD) + 5.00*(TEXT) + U10$

$B11 = -0.30$

Where ACT is the student level covariate, representing the student ACT scores, and METHOD and TEXT are class level covariates, representing teaching methods and textbook use. ACT is a continuous variable following a normal distribution with M=0 and SD=4.38. METHOD is a continuous variable following a normal distribution with M=0 and SD=0.56. TEXTBOOK is a dichotomous variable following a Bernoulli distribution with a probability p=0.66. These numbers were chosen to mimic the empirical study of the larger project.

The effect of ACT on the pre student ability and post student ability were set to 1.00 and -0.30. The METHOD and TEXTBOOK effect were set to 3.00 and 5.00 respectively. Again these numbers were chosen to resemble the effect in the empirical study, based on HLM modeling.

Random effects were set to U00 ~ N(0,2.802), U10 ~N(0,2.402), R0 ~ N(0,4.802) and R1~ N(0,1.702) so that they too mimicked the data in the empirical study.

Student response data were generated with 2PL IRT model. The item parameters were the same as in the empirical study. The item parameters are listed in appendix.

After the simulation responses were generated, the three procedures were applied to the same data to see how the explanatory variables effects were recovered.
For the simplicity of description, only the estimations of teaching method effect are compared. In the full paper, all parameters will be included

There is randomness introduced by IRT model when students respond to the items, so the estimation result is not only affected by the difference of procedures, but also by the sampling error of the students' responses. In order to even out the sampling error, 100 replications of student response data were generated and analyzed by the three different procedures.

Results

Table 1 shows the results of 100 replications of data (simulated to resemble the project data) with teaching method as a predictor, estimated using each of the three procedures. The final mean shown in Table 1 is the average of the 100 replications. The details of the 100 replications estimation will be included in an appendix to the full paper. The underlying teaching method effect was set to 3.00 in the simulated data. Note that the MCMC procedure yields a result closest to the true value of the teaching method parameter (3.00).

Table 2 shows the results of the model using project data with teaching method as a predictor, again using each of the three procedures. As of the deadline for paper submissions, the MCMC procedure for the project data was incomplete. Final results will be available for the complete paper.

Table 1.

Simulation Result of 100 Replications with Teaching Method as Predictor

Procedure	Mean, Teaching method parameter	SD	Mean SE
HLM2	3.13	0.85	0.96
HLM3	2.73	0.73	0.80
MCMC	3.03	0.78	0.92

Notes: N=100, the true teaching method effect is 3.00. In each of the 100 MCMC procedures, Heidel's test is used to make sure of the convergence of Markov Chains.

Table 2.

Models of Project Data with Teaching Method as Predictor

Procedure	Mean, Teaching method parameter	SD
HLM2	3.72	1.43
HLM3	3.06	1.13
MCMC		

In the full paper, we will include additional parameters estimated in the simulation and the real data. We will also include appendices the WinBUGS code and output from the MCMC and HLM models.

Significance of this Study

Measurement error is often ignored in multilevel modeling. But the simulation described in this paper suggests that ignoring the measurement error leads to a biased estimation of fixed effects. HLM3 growth model is a potentially useful tool for solving this problem, but simulation results reveal the flaws in this method. One possible reason for the biased estimation is that the measurement error variance estimated from IRT models is not an accurate estimation. Further investigation should explore this.

The simulation in this paper suggests that the MCMC procedure is the best tool for dealing with measurement error. Although it requires Bayesian training to apply this method, MCMC procedure can be easily carried out in the popular software WinBUGS.

Another advantage of measurement error models such as HLM3 Growth Model and MCMC is that they are better at dealing with missing data problems in practice. When a student misses pretest score or posttest score, HLM2 will have to delete the record because no gain scores can be obtained for that student if one of the test score is missing. While in HLM3 Growth Model and MCMC, the information can be used even when the pretest score or the posttest score is missing.

References

Fox, J., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling *Psychometrika, 66*(2), 271-288.

Gill, J. (2002). *Bayesian methods : a social and behavioral sciences approach*. Boca Raton, Fla.: Chapman & Hall/CRC.

Lord, F. M. (1980). *Applications of item response to theory to practical testing problems*. Hillsdale, N.J.: Erlbaum Associates.

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Publications.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). WinBUGS Version 1.2 User Manual (Version 1.2): MRC Biostatistics Unit

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software, 12*(3), 1-16.