Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 22

6-4-2019

The Impact of Equating on Detection of Treatment Effects

Youn-Jeng Choi
University of Alabama, ychoi26@ua.edu

Seohyun Kim
University of Georgia, seohyun@uga.edu

Allan S. Cohen
University of Georgia, acohen@uga.edu

Zhenqiu Lu
University of Georgia, zlu@uga.edu

The Impact of Equating on Detection of Treatment Effects

Erratum

In a previous version of this article, the third author's affiliation was incorrectly given as University of Alabama. This has been corrected to University of Georgia.

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2673
doi: 10.22237/jmasm/1559653467

בס"ד

ISSN 1538 − 9472

The Impact of Equating on Detection of Treatment Effects

Youn-Jeng Choi
University of Alabama
Tuscaloossa, AL

Seohyun Kim
University of Georgia
Athens, GA

Allan S. Cohen
University of Georgia
Athens, GA

Zenqiu Lu
University of Georgia
Athens, GA

Equating makes it possible to compare performances on different forms of a test. Three different equating methods (baseline selection, subgroup, and subscore equating) using common-item item response theory equating were examined for their impact on detection of treatment effects in multilevel models.

Keywords: Common-item IRT equating, subgroup equating, subscore equating, generalized partial credit model, mixture Rasch model, MCMC estimation

Introduction

Equating of tests is needed to compare results from different forms of a test. The impact of type of equating on detection of treatment effects is an important issue but has not been widely studied. In this study, we investigated the impact of three types of equating designs on detection of treatment effects: 1) baseline on the old form vs. baseline on the new form; 2) subgroup equating; and 3) subscore equating. Equating assumes group invariance and equity. The equity property assumes that it is a matter of indifference to each examinee which form of a test is administered (Kolen & Brennan, 2004; Lord, 1980). In an experiment, this would mean the same score should be obtained whether the base scale is from pre-test or post-test data. If selection of the base scale results in different scores, then invariance would be violated. When group invariance is violated, subgroup equating may provide a useful alternative (Cid & Spitalny, 2013; Dawber, Oh, & Wise, 2013). Subscore equating, a third method, arises mainly from the increasing demand for finer

doi: 10.22237/jmasm/1559653467 | Accepted: November 10, 2017; Published: June 4, 2019.
Correspondence: Youn-Jeng Choi, ychoi26@ua.edu

grained information (Puhan & Liang, 2011; Sinharay & Haberman, 2011). These methods were compared with respect to their impact on detection of a treatment effect.

Theoretical Framework

Common-Item IRT Equating with Different Equating Designs: Different Baseline, by Subgroup, and by Subscores

Common-item item response theory (IRT) equating often is used when more than one form per test date cannot be administered because of test security or other practical concerns. In this design, two parallel forms have a set of items in common, and different groups of examinees are administered the two forms (Kolen & Brennan, 2004). One of the simplest ways is to implement the design is to fix the item parameters for common items for both forms. When equating two or more forms, however, a baseline needs to be specified before equating. The assumption in selecting a baseline is that the choice of baseline does not affect the subsequent equating.

When the populations consist of several subgroups and each subgroup is homogeneous with respect to some characteristics such as ethnicity, gender, or item response patterns, common-item equating also can be performed based on these subgroups. For subgroup equating designs, the item and ability parameters for the common items need to be on the same scale between subgroups. When the items are scored by subscores or subsections based on content or blueprints categories, a common-item equating design can be used based on subscores (Puhan & Liang, 2011; Sinhary & Haberman, 2011).

IRT Models for Common-Item Equating

There are many designs used for collecting data for equating (Kolen & Brennan, 2004). A common-item nonequivalent group design was applied in this study. In this design, there are two nonequivalent groups. These were from the first and second year of a larger host study. All 33 items on the assessment developed for the host study were used as common items. Three different IRT models were used for common-item IRT equating methods. First, the Rasch model (Rasch, 1960) was applied to compare the effect of different baselines. Second, subgroup equating was done using latent classes estimated by a mixture Rasch model. Third, subscore equating was done using item parameters estimated from a generalized partial credit model. All three IRT models used the item centering method with the sum of item

difficulty parameter estimates equals to zero to solve the identification problem. This made it possible to put all ability and item difficulty parameters on the same scale.

The Rasch Model

The one-parameter logistic model (1PLM) has two parameters to model the relationship between abilities and responses, item difficulty (*b*) and item discrimination (*a*). The 1PLM assumes that the discrimination parameters across items are equal (i.e., $a_i = a^*$ for all i). The Rasch model (RM) is a special case of the 1PLM with all discrimination parameters fixed equal to one (Rasch, 1960). The probability that examinee j correctly answers item i (i.e., the probability that $y_{ij} = 1$) is assumed to have the following form:

$$P_{ij} = \mathrm{P}\left(y_{ij} = 1 \mid \theta_j, b_i\right) = \frac{\exp\left[a^*\left(\theta_j - b_i\right)\right]}{1 + \exp\left[a^*\left(\theta_j - b_i\right)\right]}, \tag{1}$$

where a^* are the same for all i under a 1PLM, and a^* equals one for all i under the RM. The item centering constraint ($\Sigma b_i = 0$) was used to solve the identification problem. Equation (1) is equivalently written

$$\mathrm{logit}\left(P_{ij}\right) = a^*\left(\theta_j - b_i\right), \tag{2}$$

where

$$\mathrm{logit}\left(x\right) = \log\left(\frac{x}{1-x}\right).$$

The Mixture Rasch Model

The mixture Rasch model (MRM; Rost, 1990) assumes that an examinee population is composed of a fixed number of discrete latent classes of examinees (Cohen, Wollack, Bolt, & Mroch, 2002). Each class has unique ability and item parameters. All examinees who belong to a given latent class are assumed to be homogeneous on the factor(s) that caused the latent class to form but share the same item parameter information across the latent class. The MRM in equation (3) associates a class membership parameter, g, with each examinee. This determines

the relative difficulty of the items for that examinee. Additionally, g also determines the latent ability parameter, θ_j. This, in turn, determines the number of correct answers on the test. The probability of a correct response in the MRM is written as

$$P_{ij} = \mathrm{P}\left(y_{ij} = 1 \mid \theta_j, b_{ig}\right) = \sum_{g=1}^{G} \pi_g \frac{\exp\left(\theta_j - b_{ig}\right)}{1 + \exp\left(\theta_j - b_{ig}\right)}, \tag{3}$$

where g is an index for the latent class, $g = 1,\ldots, G$; $j = 1,\ldots, N$ examinees, θ_j is the latent ability of an examinee j, π_g is the proportion of examinees for each class, and b_{ig} is the Rasch difficulty parameter of item i for class g. The indeterminacy constraint for item centering $\Sigma b_{ig} = 0$ for class g was used.

The Generalized Partial Credit Model

The generalized partial credit mode (GPCM; Muraki, 1997) is an extension of Masters' partial credit model (PCM; Masters & Wright, 1997) in which the assumption of uniform discrimination for all items is relaxed:

$$\mathrm{P}_{ih}\left(\theta_j\right) = \frac{\exp\left[\sum_{v=1}^{h} Z_{iv}\left(\theta_j\right)\right]}{\sum_{c=1}^{m_i} \exp\left[\sum_{v=1}^{c} Z_{iv}\left(\theta_j\right)\right]} \tag{4}$$

and

$$Z_{ih}\left(\theta_j\right) = a_i\left(\theta_j - b_{ih}\right) = a_i\left(\theta_j - b_i + d_h\right) \tag{5}$$

where $Z_{ih}(\theta_j)$ is the logit and $\Sigma \mathrm{P}_{ih}(\theta_j) = 1$, a_i is a slope parameter for item i and fixed to one in this study, b_{ih} is an item-category parameter for item i and hth category, b_i is an item-location parameter for item i, d_h is a category parameter for the h^{th} category of item i, and m_i is the number of response categories. To eliminate indeterminacy, b_{i1} was arbitrarily fixed to zero. The category parameters also have the following identification constraint:

$$\sum_{h=2}^{m_i} d_h = 0. \tag{6}$$

Detection of Treatment Effects in the Multilevel Model

A multilevel analysis using proc mixed SAS procedure with the maximum likelihood estimation method was performed to detect treatment effects for each of the equating methods after obtaining ability parameters (θs) using three IRT models. Post-test ability parameters (θs) estimated under each of the three equating methods were used as dependent variables. Student level manifest variables included gender, ethnicity, and language used for reading (i.e., English, Spanish, or both). Teacher level variables included level of project participation and years of teaching experience.

After comparing the six possible nested models using the likelihood ratio test, the model below was determined to be the best-fit:

Level 1 (Student level):

$$\theta_{ij} = \beta_{0j} + \beta_{1j}\text{Gender}_{ij} + \beta_{2j}\text{Hispanic}_{ij} + \beta_{3j}\text{ReadE}_{ij} + \beta_{4j}\text{ReadS}_{ij} + r_{ij}, \quad (7)$$

where $r_{ij} \sim \text{N}\left(0, \sigma^2\right)$

Level 2 (Teacher level):

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}\left(\text{Teaching}_j - 10\right) + \gamma_{02}\text{School}_j + \gamma_{03}\left(\text{Engagement}_j - 6\right) + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}\text{School}_j + \gamma_{22}\left(\text{Engagement}_j - 6\right) \quad (8) \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}\text{School}_j \\
\beta_{4j} &= \gamma_{40} + \gamma_{41}\text{School}_j,
\end{aligned}$$

where $u_{0j} \sim \text{N}\left(0, \tau_{00}^2\right)$

In equations (7) and (8), θ_{ij}, the ability parameter for student i indicates to teacher j. The variables in Level 1 were used as student-level covariates; gender was coded as 0 for female and 1 for male and Hispanic was coded as 0 for non-Hispanic and 1 for Hispanic. ReadE and ReadS are dummy coded variables. When a student read only in English, ReadE was coded as 1 and ReadS was coded as 0. When a student read only in Spanish, ReadE was coded as 0 and 1 for ReadS. When a student read in both English and Spanish, ReadE and ReadS were both coded 0 as the reference category. For the second level, the control school was coded as 0 and the treatment schools were coded as 1. Teaching was coded as the total number of years spent teaching. Teaching was centered at the median value of 10.

Engagement was coded as the number of times the teacher participated in the activities of the larger host study. Engagement was centered at the median value of 6.

Methodology

Data

Data were from the host study entitled "Language-rich Inquiry Science for English Language Learners" (LISELL). The host study was designed to improve pedagogy for teaching middle grade science (Buxton et al., 2013). There were six constructed response questions measuring four subscores: scientific inquiry, everyday language, academic language, and science content. The first two items measured use of variables, the next two items measured understanding of hypothesis, observation, and evidence, and the last two items measured cause and effect relationships. The test for the host study had two forms, one for the pretest and a second for the posttest to prevent memory effects from intruding on the data. In this study, we used only posttest data for two years and one form was selected for analysis. All six constructed response questions consist of thirty-three subquestions. Each subquestion was scored by partial credit scores. For RM and MRM, all thirty-three subquestions were dichotomously recoded: all partial credits were recoded as correct responses.

Exploratory and (when needed) confirmatory factor analyses were applied to evaluate the validity of the assessment of the dichotomously recoded written responses. Unweighted least squares extraction and direct oblimin rotation methods were used for both first- and second-year posttests. Six factors were extracted for both datasets: Factor 1 consisted primarily of Questions 1 & 2, Factor 2 consisted of Question 3, Factor 3 consisted of Question 4, Factor 4 consisted of Question 5, Factor 5 consisted of Question 6, and Factor 6 consisted of use of everyday language. The conclusion from this result was that the assessment was valid for testing understanding of science inquiry processes.

Two kinds of reliability analysis were used: inter-rater reliability using Cohen's Kappa and Cronbach's alpha for evaluating internal consistency. A random 10% sample of the tests was selected and re-scored. Inter-rater agreement (Hayes & Hatch, 1999) for this analysis indicated good consistency, .65 for the first year of the host study and .70 for the second year. Cronbach's alphas for all items for first and second years were both .91. Thus, the posttest for both years showed good reliabilities.

Table 1. Five equating designs

Design	IRT Model	Baseline (Year)	Equating method	Number of items	Number of anchors	Number of latent classes	Item type
1	RM	1st	CI	33	33	1	Dicho
2	RM	2nd	CI	33	33	1	Dicho
3	MRM	2nd	CI	33	33	4	Dicho
4	GPCM	2nd	CI	6	6	1	SbyI
5	GPCM	2nd	CI	6	6	1	SbyC

Note: CI = common-item IRT equating, Dicho = dichotomously recoded item, SbyI = subscore by item, SbyC = subscore by category

Participants

During the 2011-2012 and 2012-2013 school years, five middle schools from three school districts in a Southeastern state were recruited for the LISELL study, i.e., the larger host study. Nineteen teachers participated for the first year and 21 teachers participated for the second year across grades 6 to 8. The science teachers all were invited to participate, and those who volunteered became participants in this host study. The frequency of participation in the host study was counted. Teaching experience ranged from 3 years to 30 years. There were 1,635 students in the five participating schools: 730 students participated in the first year and 905 students participated in the second year. None of the students took the posttests for both the first and second years.

One school was selected as a control school. That school did not receive any LISELL teacher professional development. The other four schools were assigned to the treatment condition. There were 172 students in the control school and 1,435 students in the treatment school.

Results

Equating Designs

Three types of common-item equating analyses were used with each of the IRT models. The five equating designs are shown in Table 1. The RM design was used to compare the effects depending on selection of the baseline. There were two baselines considered: the posttest scale for the first year and the posttest scale for the second year. The MRM was used to compare subgroup equating with one group equating. The GPCM was used to compare two different subscore equating methods: subscore by item and subscore by category.

Bayesian IRT Estimation

Estimation of model parameters was done using Markov chain Monte Carlo (MCMC) estimation as implemented in the computer software WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). Heidelberger and Welch's convergence diagnostics (1983) were used to determine the number of iterations as implemented in the coda package using R (Plummer, Best, Cowles, Vines, & Sarkar, 2012). In addition, the ratio of the standard deviation of the estimated parameter and the MC error was also used to help determine whether the MCMC chain had converged. As a rule of thumb, the MC error should not be more than 5 percent of the standard deviation of the estimated parameter. The Heidelberger and Welch convergence diagnostic was used to decide the burn-in and post-burn-in iterations for item difficulty parameters for each IRT model. For the RM, a burn-in of 3,000 iterations was found to be sufficient for convergence for all parameters; 6,000 post-burn-in iterations were used to obtain the posterior distributions. A burn-in of 3,000 iterations and 5,000 post-burn-in iterations were used for the MRM. For the GPCM, a burn-in of 1,000 iterations and 3,000 post-burn-in iterations were used.

Estimation of Equating Designs 1 & 2 (Baseline Selection)

For equating Design 1, i.e., with the first year as the baseline, first, the RM was applied to compare equating methods using two different baselines. Item difficulty parameters and ability parameters were estimated first from the first-year dataset. Then, item difficulties estimated from the first year were fixed and used to estimate ability parameters for the second-year dataset. For equating Design 2, i.e., the second year as the baseline, item difficulty parameters and ability parameters from the second-year posttest data were estimated. Then item difficulties estimated from second year were fixed and applied to estimate ability parameters for the first year (See Table 2).

Before equating, the mean of ability parameters for first year was .00 and .33 for second year. After equating using Design 1, the mean ability for the first year was .00 and .21 for the second year. When Design 2 was applied, the mean ability for first year was .11; for the second year, it was .33. Thus, there were differences in ability between Design 1 and Design 2.

Table 2. Item difficulty parameters for IRT anchor equating based on two baselines

Item No.	Description	Baseline (1st year)	Baseline (2nd year)	Difference
1	InquiryQ1	0.15	0.15	0.00
2	EVLanguageQ1	−1.69	−1.77	0.09
3	ACLanguageQ1	0.42	0.35	0.07
4	ContentQ1	0.82	1.24	−0.42
5	InquiryQ2_1	0.76	0.65	0.12
6	InquiryQ2_2	1.10	1.27	−0.16
7	InquiryQ2_3	0.77	0.67	0.10
8	EVLanguageQ2	−2.83	−2.87	0.04
9	ACLanguageQ2	0.59	0.39	0.20
10	ContentQ2	−0.16	−0.03	−0.13
11	InquiryQ3_1	−1.50	−1.60	0.10
12	InquiryQ3_2	1.11	1.07	0.03
13	EVLanguageQ3	−3.53	−3.21	−0.32
14	ACLanguageQ3	−0.05	−0.17	0.12
15	ContentQ3	0.92	0.80	0.13
16	InquiryQ4_1	−0.28	−0.43	0.15
17	InquiryQ4_2	1.44	1.33	0.11
18	InquiryQ4_3	1.99	1.91	0.08
19	EVLanguageQ4	−2.53	−2.11	−0.42
20	ACLanguageQ4	0.46	0.30	0.16
21	ContentQ4	0.79	0.91	−0.12
22	InquiryQ5_1	1.12	0.92	0.20
23	InquiryQ5_2	1.06	0.78	0.28
24	EVLanguageQ5	−2.21	−1.59	−0.62
25	ACLanguageQ5	1.25	0.77	0.49
26	ContentQ5	0.95	0.62	0.33
27	InquiryQ6_1	0.10	0.06	0.04
28	InquiryQ6_2	0.81	0.63	0.18
29	InquiryQ6_3	0.22	0.24	−0.03
30	InquiryQ6_4	0.64	0.80	−0.16
31	EVLanguageQ6	−2.69	−2.21	−0.49
32	ACLanguageQ6	−0.01	0.14	−0.16
33	ContentQ6	0.00	0.01	0.00

Table 3. Model selection indices to find the number of latent classes

Number of latent classes	AIC		BIC	
	1st year	2nd year	1st year	2nd year
1	22620	28260	22770	28420
2	20840	25810	21150	26140
3	19720	24840	20200	25340
4	19250	23740	19890	24410
5	24240	30010	25040	30840

Table 4. Item difficulty parameters of four latent classes using subgroup IRT equating

Item No.	Description	Class 1	Class 2	Class 3	Class 4
1	InquiryQ1	0.89	−0.29	−0.49	0.18
2	EVLanguageQ1	−0.35	−2.13	−2.86	−1.47
3	ACLanguageQ1	1.17	−0.05	−0.39	0.39
4	ContentQ1	2.67	1.31	−0.34	1.11
5	InquiryQ2_1	1.05	0.65	0.23	0.63
6	InquiryQ2_2	2.12	1.26	0.33	1.15
7	InquiryQ2_3	1.16	0.50	0.06	0.83
8	EVLanguageQ2	−2.70	−2.78	−3.37	−2.93
9	ACLanguageQ2	0.94	0.46	−0.38	0.51
10	ContentQ2	0.82	−0.27	−0.88	0.05
11	InquiryQ3_1	−0.38	−1.49	−2.67	−1.36
12	InquiryQ3_2	2.26	0.99	−0.18	0.97
13	EVLanguageQ3	−2.68	−3.18	−3.69	−3.27
14	ACLanguageQ3	1.07	−0.51	−1.20	−0.24
15	ContentQ3	2.05	0.74	−0.43	0.63
16	InquiryQ4_1	0.14	−0.52	−1.09	−0.37
17	InquiryQ4_2	1.85	1.79	0.53	1.30
18	InquiryQ4_3	2.54	2.29	1.19	1.67
19	EVLanguageQ4	−1.62	−2.34	−2.81	−1.81
20	ACLanguageQ4	1.00	−0.11	−0.36	0.39
21	ContentQ4	1.42	1.24	0.37	0.72
22	InquiryQ5_1	−0.52	−1.25	3.09	4.26
23	InquiryQ5_2	−1.22	−1.54	2.95	3.74
24	EVLanguageQ5	−3.17	−3.40	−1.96	−0.99
25	ACLanguageQ5	0.29	−0.61	1.25	1.90
26	ContentQ5	−2.61	−2.47	3.16	3.32
27	InquiryQ6_1	−1.17	1.87	1.26	−1.65
28	InquiryQ6_2	−0.30	3.36	2.43	−0.46
29	InquiryQ6_3	−0.97	2.32	2.23	−1.52
30	InquiryQ6_4	0.00	3.69	2.81	−0.18
31	EVLanguageQ6	−3.14	−2.52	−2.23	−4.07
32	ACLanguageQ6	0.07	1.10	0.44	−0.83
33	ContentQ6	−2.72	1.89	2.99	−2.60

Note: EV = Everyday, AC = Academic

Estimation of Equating Design 3 (Subgroups using Latent Classes)

An exploratory MRM analysis was done using the MCMC algorithm as implemented the computer code WinBUGS. The exploratory analysis was done to determine the number of latent groups in the data. Solutions for one to five latent classes were fit. The number of latent classes was determined using the Akaike's information criterion (AIC) and Bayesian information criterion (BIC) as suggested by Li, Cohen, Kim, and Cho (2009). The smaller AIC and BIC values represent

better model fit. BIC and AIC values in Table 3 indicated that a model with four latent classes was the best fit to both first- and second-year datasets.

The ability parameters from the four-group MRM were then used to obtain ability parameter estimates for Equating Design 3. First, difficulty parameters were estimated from the second-year data. These were then fixed and used to estimate the ability parameters for the first-year data. Table 4 shows the difficulty parameters for each of the four latent classes. Examinees were classified using the modes of posterior densities for group membership into one of the four latent groups detected in the exploratory analysis. Class 1 was the highest ability group, Classes 2 and 4 were average ability groups, and Class 3 was the lowest ability group. The sample sizes of the lowest group (Class 3), $N = 84$ and 315, were the largest for both years, respectively.

Results in Figure 1 clearly show that the item difficulty parameters differed across the four latent classes. Questions 5 and 6 were much easier for Class 1 (the highest ability group) than for Class 3 (the lowest ability group). This suggests that members of Class 1 had greater knowledge about cause and effect than did members of Class 3. Although members in Classes 2 and 4 had similar average abilities (see Table 5), the item difficulties were differentially difficult for Questions 5 and 6 in these two classes. There was, however, no meaningful difference in difficulty parameters between Classes 2 and 4 for Questions 1 to 4, but members of Class 2 did perform better than the members of Class 4 on Question 5.

In contrast, Class 4 had lower difficulty parameters than Class 2 for Question 6 indicating that Question 6 was easier for members of Class 4. Both Questions 5 and 6 assessed knowledge about the relationship between cause and effect. The two questions differed in that Question 5 used everyday language (e.g., using words like if and then) in the question whereas Question 6 used academic language (e.g., using words like cause and effect). Therefore, although the mean ability in Classes 2 and 4 were similar, there appears to be a difference in understanding and use of everyday language and academic language with respect to the cause and effect.

The proportions of group membership and mean ability estimates between first and second year were similar to each other using Design 3 (subgroup common-item IRT equating). There do not appear to be differences in ability for the four latent groups in the first and second year.

Cross tabulation analyses of manifest groups (e.g., school type. race, gender, language usage for reading, teaching year and teacher's engagement, etc.) and latent class were done did not indicate differences on these variables across latent classes. An example is given in Table 6.

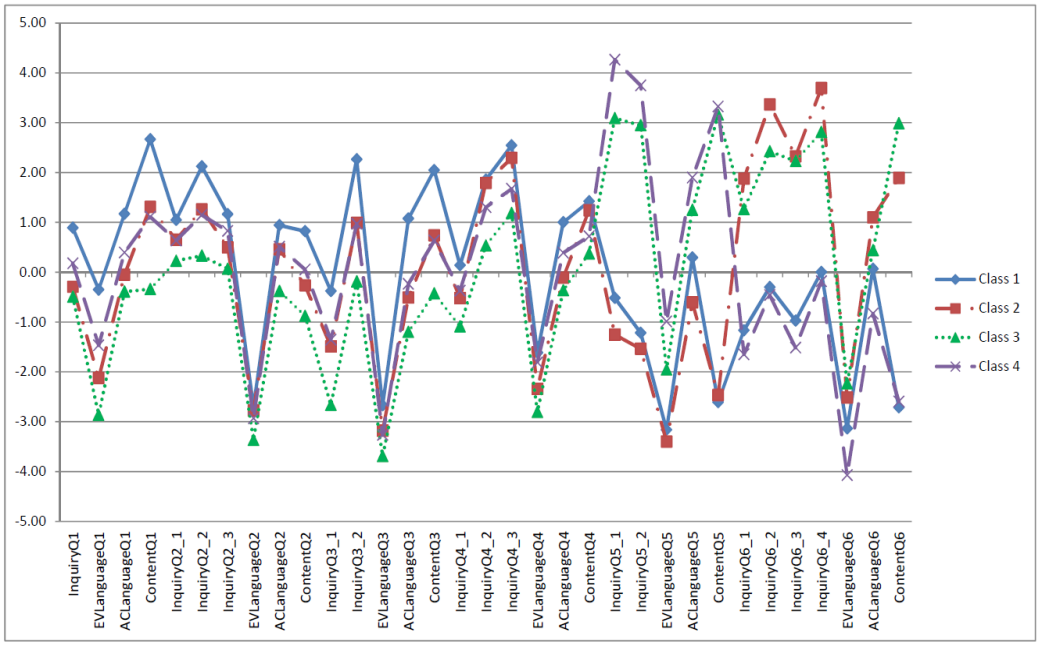

Figure 1. Item difficulty comparison for four latent classes

Table 5. Equated mean of ability parameters and latent group proportions using subgroup equating

	1st year			2nd year			Total	
Latent class	**Mean**	***n***	**%**	**Mean**	***n***	**%**	***n***	**%**
1	1.67	183	25.1	1.78	277	30.6	460	28.1
2	0.50	69	9.5	0.57	110	12.2	179	10.9
3	−1.44	284	38.9	−1.47	315	34.8	599	36.6
4	0.40	194	26.6	0.31	203	22.4	397	24.3
Total		730	100.0		905	100.0	1635	100.0

Table 6. Latent classes make-up by control and treatment school (equating design 3)

	Class 1	**Class 2**	**Class 3**	**Class 4**	**Total**
Control school	59 (34.5%)	19 (11.1%)	37 (21.6%)	56 (32.7%)	171 (100.0%)
Treatment school	375 (27.2%)	154 (11.2%)	525 (38.0%)	327 (23.7%)	1381 (100.0%)
Total	434 (28.0%)	173 (11.1%)	562 (36.2%)	383 (24.7%)	1552 (100.0%)

Estimation of Equating Designs 4 and 5 (Subscores from Six Questions and Subscores from the Six Item Categories)

Subscore equating has recently been suggested as having potential remedial and instructional benefits (Puhan & Liang, 2011; Sinhary & Haberman, 2011). The assessment used in this study had two different subscores: six questions (Questions 1 to 6) and six categories: 3 categories of scientific inquiry, one category of everyday language, one category of academic language, and one category of science content. The second year was used as the baseline for Designs 4 and 5.

The parameters of the six subscores were estimated using the GPCM. Tables 7 and 8 present the item location (b) and category (τ) parameters for the GPCM for subscores estimated by item and for subscores estimated by category. IRT common-item equating was used for equating both subscores.

Table 7. Item location (b) and category (τ) parameters for six subscores by question

Item parameter	Q1	Q2	Q3	Q4	Q5	Q6
b	0.62	0.50	0.13	0.71	0.62	0.48
τ_0	0.00	0.00	0.00	0.00	0.00	0.00
τ_1	0.47	1.60	0.96	0.81	0.88	0.99
τ_2	0.28	0.06	0.75	0.73	−1.17	−0.56
τ_3	0.39	−0.07	−0.04	0.13	−0.78	−0.48
τ_4	−1.14	−0.23	−1.87	−0.64	0.87	−0.01
τ_5		−0.67	0.20	−0.12	0.20	0.15
τ_6		−0.69		−0.91		−0.45
τ_7						0.36

Table 8. Item location (b) and category (τ) parameters for six subscores by category

Item	Inquiry			Language		
parameter	Qs 1 & 2	Qs 3 & 4	Qs 5 & 6	Everyday	Academic	Content
b	0.11	−0.03	−0.02	−1.91	−0.21	0.14
τ0	0.00	0.00	0.00	0.00	0.00	0.00
τ1	0.84	1.87	−0.25	−0.06	1.14	1.49
τ2	0.24	0.76	0.93	0.88	0.79	0.97
τ3	−0.17	0.11	−0.64	−0.13	0.30	0.67
τ4	−0.90	−0.92	0.87	0.24	−0.46	−0.16
τ5		−1.82	−1.14	−0.64	−0.58	−1.18
τ6			0.22	−0.28	−1.19	−1.80

Table 9. Descriptive analysis for Level 1 and 2 variables in the multilevel modeling

		Control school				Treatment school			
	Variable	***n***	**%**	**Mean**	**SD**	***n***	**%**	**Mean**	**SD**
	Teaching	169		16.88	11.54	1347		13.25	*8.11*
	Engagement	169		0.21	0.41	1347		7.89	5.92
Estimates	θ_1	169		0.62	1.01	1347		0.13	1.18
	θ_2	169		0.72	0.99	1347		0.24	1.15
	θ_3	169		0.77	1.33	1347		0.10	1.54
	θ_4	169		0.88	0.56	1347		0.61	0.66
	θ_5	169		0.08	0.80	1347		−0.31	0.91
Gender	Female	84	49.7			670	49.7		
	Male	85	50.3			677	50.3		
Race	Non-Hispanic	92	54.4			732	54.3		
	Hispanic	77	45.6			615	45.7		
Reading	English	137	81.1			1012	75.1		
	Spanish	2	1.2			31	2.3		
	Both	30	17.8			304	22.6		

Multilevel Modeling to Detect Treatment Effects

Descriptive Analysis. There were 668 students and 19 teachers for year 1 and 848 students and 21 teachers for year 2 following listwise deletion. Teachers assigned to the treatment schools had eight times more engagement with the LISELL project than teachers from the control school. Teachers from the control school had 3.5 years more of teaching experience on average than teachers from the treatment schools.

There was no difference between the control and treatment schools on student background information such as gender, race, and language used for reading. Five different ability parameters were estimated from the five different equating designs. For the control school, mean ability parameters (ranging from $M(\theta_1) = 0.62$ to $M(\theta_4) = 0.88$) appeared to be similar except for the mean ability ($M(\theta_5) = 0.08$) from Design 5 (see Table 9).

For treatment schools, the three mean ability parameters (ranging from $M(\theta_3) = 0.10$ to $M(\theta_2) = 0.24$) estimated using Designs 1 to 3 appeared to be similar. The mean ability parameter for Design 4 $M(\theta_4) = 0.61$ for the treatment schools using subscore equating by question was higher and the mean ability parameter $M(\theta_5) = -0.31$ for treatment schools using subscore equating by category had the lowest values. Overall, the control school had higher mean ability than the

treatment schools for on all five designs. Three equating designs, Designs 1 to 3, using two different baselines with the RM and using four subgroups detected with the MRM produced similar mean ability parameter estimates. The mean ability parameter from subscore equating by category was lowest and that from subscore equating by item was highest among all five ability parameter estimates (see Table 9).

Treatment Effect Interpretation from the Multilevel Models. Five ability parameters were applied to the multilevel model as a dependent variable. Ability estimates were obtained from each of the five equating designs. Ability means for the RM were estimated using common-item equating for Designs 1 and 2. The mean ability estimate for Design 3 (θ_3) was estimated using subgroup based on latent classes. Designs 4 and 5 using the GPCM provided mean ability estimates (θ_4 and θ_5) for the multilevel model described in equations (7) and (8). The ICC values for five equating designs were .12-.13. It tells us that about 12-13% of the variability in outcome difference occurs between Level 2 units.

Results in Table 10 compare fixed and random effects on the same multilevel models for the five equating designs. The school (γ_{02}) and engagement (γ_{03}) variables, which measure the treatment effect, were not significant for any of the equating designs. The interaction between engagement and Hispanic (γ_{22}) was significant at the .05 level for Designs 1, 2, 3, and 5, although the coefficient values were very small (ranging from .01 to .03). The intercepts (γ_{00}) from Designs 1 to 4 looked similar but that for Design 5 was smaller.

The coefficients in the second and third columns of Table 10 suggest there was no difference under the two different baseline selections. When Designs 2 and 3 were used, coefficients related to reading in Spanish (γ_{40} and γ_{41}) were not significantly different. There were differences on the random effect. This may indicate that ability parameters estimated from Design 3 were less well explained by the proposed model than ability estimates from Design 2. As noted above, manifest variables were not related to characteristics of the four latent classes (see Table 6 and Figure 1). The fixed effects from Designs 4 and 5 were similar but the random effect of Design 5 was larger than that of Design 4. This suggests that ability parameters estimated from Design 5 were less well explained by the proposed model than ability estimates from Design 4.

Table 10. Treatment effect comparisons on the inquiry science assessment using five equating designs

		Design 1 (Baseline: 1st year)			Design 2 (Baseline: 2nd year)			Design 3 (Subgroup)			Design 4 (Subscore: Item)			Design 5 (Subscore: Category)		
Effect	**Variables**	***B***	**(SE)**	***p***	***B***	**(SE)**	***p***	***B***	**(SE)**	***p***	***B***	**(SE)**	***p***	***B***	**(SE)**	***p***
	Intercept (γ_{00})	0.65	−0.40	0.12	0.75	−0.40	0.07	0.75	−0.53	0.17	0.84	−0.22	0.00*	0.03	−0.30	0.93
Fixed	Teaching (γ_{01})	0.01	−0.01	0.53	0.01	−0.01	0.55	0.01	−0.01	0.53	0.00	−0.01	0.43	0.01	−0.01	0.51
	School (γ_{02})	−0.27	−0.42	0.52	−0.27	−0.42	0.51	−0.31	−0.55	0.57	−0.08	−0.23	0.71	−0.14	−0.32	0.67
	Engagement (γ_{03})	0.00	−0.01	0.94	0.00	−0.01	0.81	−0.01	−0.02	0.43	−0.01	−0.01	0.27	−0.01	−0.01	0.35
	Gender (γ_{10})	−0.11	−0.06	0.04*	−0.11	−0.05	0.04*	−0.11	−0.07	0.15	−0.07	−0.03	0.02*	−0.09	−0.04	0.04*
	Hispanic (γ_{20})	−0.06	−0.19	0.76	−0.05	−0.18	0.78	−0.07	−0.25	0.79	−0.04	−0.10	0.67	−0.05	−0.15	0.73
	Hispanic*School (γ_{21})	−0.30	−0.21	0.15	−0.30	−0.20	0.14	−0.42	−0.27	0.13	−0.14	−0.12	0.21	−0.22	−0.16	0.17
	Hispanic*Engagement (γ_{22})	0.02	−0.01	0.03*	0.02	−0.01	0.03*	0.03	−0.01	0.03*	0.01	−0.01	0.06	0.02	−0.01	0.03*
	ReadE (γ_{30})	0.12	−0.24	0.61	0.13	−0.23	0.58	0.10	−0.31	0.76	0.08	−0.13	0.53	0.11	−0.18	0.54
	ReadE*School (γ_{31})	−0.15	−0.25	0.56	−0.15	−0.24	0.53	−0.13	−0.33	0.68	−0.09	−0.14	0.50	−0.13	−0.19	0.52
	ReadS (γ_{40})	−1.40	−0.78	0.07	−1.34	−0.76	0.08	−1.84	−1.02	0.07	−1.00	−0.44	0.02*	−1.04	−0.60	0.09
	ReadS*School (γ_{41})	0.51	−0.81	0.53	0.48	−0.79	0.54	0.76	−1.06	0.47	0.41	−0.45	0.36	0.39	−0.62	0.54
Random	UN(1,1) (τ^2_{00})	0.17	−0.06		0.17	−0.06		0.29	−0.09		0.05	−0.02		0.09	−0.03	
	Residual (σ^2)	1.14	−0.04		1.08	−0.04		1.96	−0.07		0.35	−0.01		0.68	−0.03	

Note: * < .05, Teaching and Engagement variables were centered to their median values

Discussion

Three different equating methods (baseline selection, subgroup, and subscore) were examined for their impact on the detection of treatment effects in an empirical dataset. Common-item IRT equating was used with all three equating methods. Results indicated that selection of a baseline did not affect the equated abilities. In addition, subgroup equating using the latent classes extracted by the MRM provided similar equated mean ability parameters using one group equating. However, the use of the different subscores produced different equated ability scores.

Equated ability parameters were similar for Designs 1 to 3: baseline comparison (year 1 vs. year 2) and one group equating using RM vs. latent subgroups detected using the MRM, respectively. Subscore equating by item tended to overestimate ability parameters. On the other hand, subscore equating by category tended to underestimate ability parameters (see Table 9). There was no significant impact of equating on detection of treatment effects. The coefficients to measure effects for school and engagement variables were close to zero.

Posttest data were used for estimating the item and ability parameters. This was primarily because knowledge of science inquiry practices was lower on the pretest. On the pretest, students had already had the instructional intervention in the treatment condition. These item difficulty parameter estimates were then used for estimating ability under each of the design conditions. The GPCM was used to estimate the item and ability parameters as a way of estimating these parameters given the testlet structure of the test. It would be interesting to see whether including a testlet model would provide different estimates than were obtained with the GPCM.

For subgroup equating, the manifest variables were used to measure the treatment effect did not appear to explain the characteristics of the four latent classes. These manifest variables are common in school-based instructional intervention research. It would be interesting to examine whether manifest variables about students and teachers that were more closely related to the instructional intervention might be related to latent class membership and, possibly, to the equating effect.

References

Buxton, C. A., Allexsaht-Snider, M., Suriel, R., Kayumova, S., Choi, Y.-J., Bouton, B., & Baker, M. (2013). Using educative assessments to support science teaching for middle school English-language learners. *Journal of Science Teacher Education, 24*(2), 347-366. doi: 10.1007/s10972-012-9329-5

Cid, J., & Spitalny, I. (2013, April). *Investigating the effect of language on the invariance of equating functions in paper-and-pencil and computer-based tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cohen, A. S., Wollack, J. A., Bolt, D. M., & Mroch, A. A. (2002, April). *A mixture Rasch model analysis of test speededness*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Dawber, T., Oh, H.-J., & Wise, M. (2013, April). *Estimation of population invariance in true-score equating for special education and non-special education student groups*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*(3), 354-367. doi: 10.1177/0741088399016003004

Heidelberger, P., & Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research, 31*(6), 1109-1144. doi: 10.1287/opre.31.6.1109

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Newbury Park, NY: Springer. doi: 10.1007/978-1-4757-4310-4

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*(5), 353-373. doi: 10.1177/0146621608326422

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

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325-337. doi: 10.1023/a:1008929526011

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response*

theory (pp. 101-122). New York, NY: Springer. doi: 10.1007/978-1-4757-2691-6_6

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York, NY: Springer. doi: 10.1007/978-1-4757-2691-6_9

Plummer, M., Best, N., Cowles, K., Vines, K., & Sarkar, D. (2012). coda: Output analysis and diagnostics for MCMC [R software package]. Retrieved from https://cran.r-project.org/package=coda

Puhan, G., & Liang, L. (2011). Equating subscores using total scaled scores as an anchor (Report No. ETS RR-11-07). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2011.tb02243.x

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271-282. doi: 10.1177/014662169001400305

Sinharay, S., & Haberman, S. J. (2011). Equating of augmented subscores. *Journal of Educational Measurement, 48*(2), 122-145. doi: 10.1111/j.1745-3984.2011.00137.x