ED 478 422 TM 035 106

AUTHOR	Michaelides, Michalis P.
TITLE	Sensitivity of IRT Equating to the Behavior of Test Equating Items.
PUB DATE	2003-04-00
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Achievement Gains; *Equated Scores; *High School Students; High Schools; *Item Response Theory; *Test Items
IDENTIFIERS	Delta Plot Method

ABSTRACT

The delta-plot method is used to identify which common items in a common item nonequivalent groups design for test equating show large changes in their p-values across administrations. Outliers in that plot denote differential item behavior and are candidates for exclusion from the common item pool. This study investigated whether keeping or discarding those outliers has an effect on equating transformations and equated aggregates of the score distributions. Two consecutive assignments from four statewide programs were analyzed, with the item response theory (IRT) mean/sigma method used for equating the year 2 to the Year 1 tests. Samples ranged from 7,128 to 17,737 high school students. Effects are more pronounced on the average gains from one year to the next than on individual scores and slightly more so when a three-parameter logistic versus a one-parameter logistic model is used for test calibration. (Contains 1 figure, 7 tables, and 11 references.) (Author/SLD)

Sensitivity of IRT equating to the behavior of test equating items

Michalis P. Michaelides

Stanford University

Paper presented at the AERA Annual Meeting in Chicago Il., April 2003

ABSTRACT

The delta-plot method is used to identify which common items in a common item nonequivalent groups design for test equating show large changes in their p-values across administrations. Outliers in that plot denote differential item behavior and are candidates for exclusion from the common item pool. This study investigates whether keeping or discarding those outliers has an effect on equating transformations and equated aggregates of the score distributions. Two consecutive assessments from four statewide programs are analyzed, with the IRT mean/sigma method used for equating the Year 2 to the Year 1 tests. Effects are more pronounced on the average gains from one year to the next than on individual scores and slightly more so when a 3-parameter logistic versus a 1-parameter logistic model is used for test calibration.

Introduction

Large-scale testing programs are administered over multiple occasions. If a single test form is handed out repeatedly, concerns about the security and overexposure of the content will arise; to overcome this problem, testing programs develop alternate test forms, according to a blueprint. When examinees take different forms, it does not follow that the scores they obtain are comparable, because forms differ in various respects, such as their degree of difficulty. Direct comparisons between examinees are not fair, unless their scores are adjusted to take account of these differences. Equating is the statistical process that establishes comparability between alternate forms of a test built to the same content and statistical specifications by placing scores on a common scale; thus, allowing interchangeable use of scores on these forms (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME], 1999).

The common-items nonequivalent groups design (Kolen and Brennan, 1995) is a commonly used design for constructing, administering and equating test forms over different administrations of a testing program. In this design, a subset of items is embedded in any two, or more, to-be-equated forms to provide a standard on which the performance of groups responding to different forms is compared. These common items constitute the equating – also called linking or anchor – item pool and serve to generate an equating function that places the scores earned on different forms on the same scale.

This study investigates whether and to what extent decisions as to which items to treat as common impact the equating transformations, and the resulting score distribution summaries. Because common items need to have good psychometric properties, they are scrutinized for the consistency of their behavior across administrations of assessments. The delta-plot procedure is a method that identifies which items demonstrate large deviations in their difficulty indices between any two forms or administrations. The effects on the equating transformation and the equated score aggregates are examined, when all common items are used vis-à-vis when misbehaving items identified by the delta-plot method have been discarded from the common item pool. In addition, the 3- versus the 1- parameter Item Response Theory (IRT) model for calibration is compared to test whether the effect of including or excluding items with large change in their difficulty indices differs according to the model employed.

Theoretical Framework

The behavior of equating items must be consistent across two forms or administrations before they are considered appropriate for the equating process. Given that the statistical properties of an item are relatively stable, any differential examinee performance on that item can be ascribed to changes in the proficiency of the examinee groups that respond to it on different occasions. If an equating item demonstrates large change in its difficulty index, for example, it calls for inspection to try and determine what caused the change. Such an item is likely to be discarded from the equating item pool and treated as a regular, non-common item (Kolen and Brennan, 1995).

There exists ample evidence in the literature demonstrating that item characteristics change when obtained from different groups. This is true both with classical and IRT statistics. In particular, in IRT the assumption is that item parameters are invariant irrespective of the group of examinees used to estimate them. Similarly, ability estimates are supposed to be invariant across groups of items taken by examinees (e.g. Hambleton, Swaminathan, and Rogers, 1991; Lord, 1980). Parameter invariance is a very useful property, but depends to a large extent on how well the IRT model's assumptions, unidimensionality, in particular, hold (Miller and Linn, 1988). Findings have been consistent both that unidimensionality is often not met and that item parameters are not invariant across different administrations. For example, differential instruction and different curricular emphases introduce additional dimensions and elicit differential examinee performance (e.g. Miller and Linn, 1988; Bock, Muraki and Pfeiffenberger, 1988). Such situations will result in items having differential difficulty for different groups of examinees. Apart from content reasons, item parameter change has been detected due to context effects such as change in the positioning, as well as changes in the wording of items (e.g. Yen, 1980).

IRT or classical item statistics may be used to examine whether items are functioning differentially for groups taking different test forms with embedded common items (Kolen and Brennan, 1995). A simple and comprehensible method for studying the item-by-group interaction is the delta-plot procedure (Angoff, 1972). It can be applied to examine the volatility of item p-values, the classical test theory difficulty indices. It is widely used because it is practical, easy to implement, does not require IRT calibrations (which would be the case if IRT parameters were compared) and because it provides prima-facie evidence regarding large changes in item difficulties across administrations.

In the delta-plot procedure, p, the proportion correct of the common item is transformed to the delta metric through a linear transformation of the inverse normal equivalent (Dorans and Holland, 1993):

$$\Delta = 13 - 4\ \{\Phi^{-1}(p)\} \qquad [1]$$

The delta metric has a mean of 13 and a standard deviation of 4 and larger values correspond to more difficult items, as opposed to the proportion correct scale, which is bounded between 0 and 1 with easier items having higher values than more difficult ones.

Two groups respond to the same items, the item p-values for each group are estimated, transformed to the delta metric, and plotted on a scatter plot. Each point on the plot corresponds to an item with the delta value for one group on the horizontal axis and the delta value for the other on the vertical axis. Outliers denote items that are functioning differentially for the two groups with respect to the level of difficulty. A handy rule to determine which items are outliers draws a "best-fit" line to the points and calculates the perpendicular distances of each point to the line. The fitted line is chosen so as to minimize the sum of squared perpendicular distances (and not the sum of squared vertical distances as in ordinary least squares regression) of the points to the line. Any point lying more than 3 standard deviations of the distances away from the line is a candidate for exclusion from the common item pool. Such items call for inspection to determine plausible causes for the differential performance in the two groups.

The purpose of this study is to examine whether the exclusion of common items, which elicit differential examinee performance across administrations, from further consideration in the equating process makes a substantial difference in the resulting equated score distributions. If it does, then the discarding of common items might deserve more consideration before undertaking such an action. And a method other than the delta-plot procedure, a method that is less arbitrary and more consistent with the measurement models used in equating, might be preferable.

Methods and Data Sources

Data from four statewide assessment programs from three states were analyzed. For each of the four assessments, there were data from two successive annual administrations: Year 1 and Year 2, thus allowing equating to be carried out for each one. Table 1 gives characteristics of the assessments. Various grades participated in these assessments in the areas of mathematics, science, and social studies. The populations tested constituted the annual cohort of students graduating in the respective states and are relatively large, ranging from 7128 to 17371.

TABLE 1

Information for the assessment data analyzed

Subject	Grade	State	Year 1 number of examinees	Year 2 number of examinees
Mathematics	8	1	7258	7128
Science	11	2	14244	14565
Social Studies	6	3	17126	17371
Science	6	3	17128	17371

Each assessment consisted of multiple forms. A large number of both dichotomously and polytomously scored items were arranged in the various forms in a matrix sampling design. Each examinee responded to only a subset of those items. The common item pools consisted of both dichotomous and polytomous items. The total number of items and the number of common items in each test are shown in Table 2.

TABLE 2

Item information for each assessment

Assessment	Total number of items (Year1/Year 2)	Number of common items
Mathematics 8	139/137	44
Science 11	126/138	45
Social Studies 6	124/95	56
Science 6	123/95	50

To carry out the delta-plot procedure the Year 1 and Year 2 p-values for each common item were calculated: the proportion correct for the dichotomous items and the mean score over the maximum possible score for the polytomous items. For each pair of consecutive assessments, e.g. the Mathematics grade 8 in Year 1 and Year 2, a delta plot was constructed plotting the Year 1 versus the Year 2 p-values, which were transformed into the delta metric with equation 1. A line was fitted in each plot to identify outlying points, as described in the previous section. The slope for the "best-fit" line was estimated by the ratio of standard deviations of the transformed p-values and the intercept was determined such that the line passes through the point where the abscissa is the mean of the transformed p-values for Year 1 and the ordinate is the mean of the transformed p-values for Year 2. Any points lying more than 3 standard deviations of all distances away from that line were signified as outliers.

Prior to estimating equating transformations, each test was calibrated separately using PARSCALE 3.0 software (Muraki & Bock, 1997) with a 3- parameter logistic (3PL) IRT model.

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{D\alpha_i(\theta\text{-}b_i)}}{1 + e^{D\alpha_i(\theta\text{-}b_i)}} \qquad [2]$$

$P_i(\theta)$ is the probability that an examinee with ability level θ answers item i correctly. D is constant and when is set equal to 1.7 the logistic model is very close to the normal ogive (Lord, 1980). For each item, the model provides three parameter estimates α, b, and c; the discrimination, difficulty, and pseudoguessing parameters respectively. The assessments included polytomous as well as dichotomous items and all were scaled together. For the polytomous items, a graded response model (Samejima, 1969) was fitted. The logistic form of the graded response model is (Muraki and Bock, 1997):

$$P_{ij}(\theta) = \frac{e^{D\alpha_i(\theta - b_i + d_j)}}{1 + e^{D\alpha_i(\theta - b_j + d_j)}} - \frac{e^{D\alpha_i(\theta - b_i + d_{j+1})}}{1 + e^{D\alpha_i(\theta - b_j + d_{j+1})}} \qquad [3]$$

where $P_{ij}(\theta)$ is the probability that an examinee with ability level θ obtains a score j (j=0,…,m) on item i with m+1 scoring categories, and d_j is the category parameter (b_i-d_j is also referred to as category threshold parameter.)

Often the calibrations in PARSCALE 3.0 did not converge. Some items that had large standard errors in their IRT b parameters, or for which very few examinees had selected one of the responses were skipped from calibration until the procedure converged. In the data sets presented in this study none of the skipped items was a common item.

The assessments came in multiple forms every year with certain items embedded in all those forms – in addition to the common items embedded across years for equating cohort scores. Concurrent calibration of all forms for a given year automatically places them on a single scale.

The classical test theory difficulty values determine which common items to keep for equating purposes through the delta-plot. The IRT b values for the common items serve to generate the equating transformation. One IRT-moments method for equating is the mean/sigma method (Kolen and Brennan, 1995). It provides a transformation function of the form

$$\theta_j = A\, \theta_i + B \qquad [4]$$

where θ_i is the ability scale for year i and A and B are constants. IRT ability estimates and IRT b parameters are on the same scale, thus the same constants can place two sets of b parameters on the same scale in a similar fashion:

$$b_j = A\, b_i + B \qquad [5]$$

The constants can be estimated using the means and standard deviations of the IRT b parameters of the common items as follows:

$$A = \frac{s(b_j)}{s(b_i)}, \text{ and} \qquad B = \bar{b}_j - A\, \bar{b}_i \qquad [6a, 6b]$$

Equation 4 provides a transformation, which places the Year 2 ability estimates on the same scale as the Year 1 ability estimates. For each pair of assessments (Year 1 and Year 2) the equating constants A and B are estimated in two ways: (a) using the IRT b parameters of all common items embedded in the assessments, and (b) using the IRT b parameters of the common items that were not signified by the delta-plot procedure as outliers. Then each one of these transformations is applied on the Year 2 distribution aggregates to rescale them onto the same scale as the Year 1 distribution aggregates, and render cross-sectional scores comparable.

Finally, the methodology just described is carried out when a 1-parameter logistic (1PL) model is fitted:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad [7]$$

by fixing the α parameters at 1 and excluding c from the model. The graded response model (equation 3 with the α parameters fixed at 1) was fitted to the polytomous data. Results between the two IRT models are compared.

Results

The first step in the analysis was the construction of delta plots for each Year 1 versus Year 2 assessment. The delta plot for the Mathematics grade 8 assessment is shown in Figure 1. Each point represents the p-values of a common item in Year 1 and Year 2 transformed to the delta metric. Outliers lying more than three standard deviations of the distances of the points to the fitted line are marked as solid squares. The number of outlying, and thus discarded by the delta-plot, common items is shown on Table 3. In all cases, outliers were less than 7% of all common items.

Having identified which items function differentially for the two annual cohorts, two sets of common items can be considered: (1) all common items, and (2) the common items that were not outliers on the delta plots. Table 3 also lists the slope A and intercept B for each of the four assessments equating transformations derived with each of the two different sets of common items: first, when all the common items' b parameters were used to derive the constants, and second, when only the items that were not identified by the delta-plot as misbehaving.

FIGURE 1

Delta plot for the Mathematics grade 8 assessment

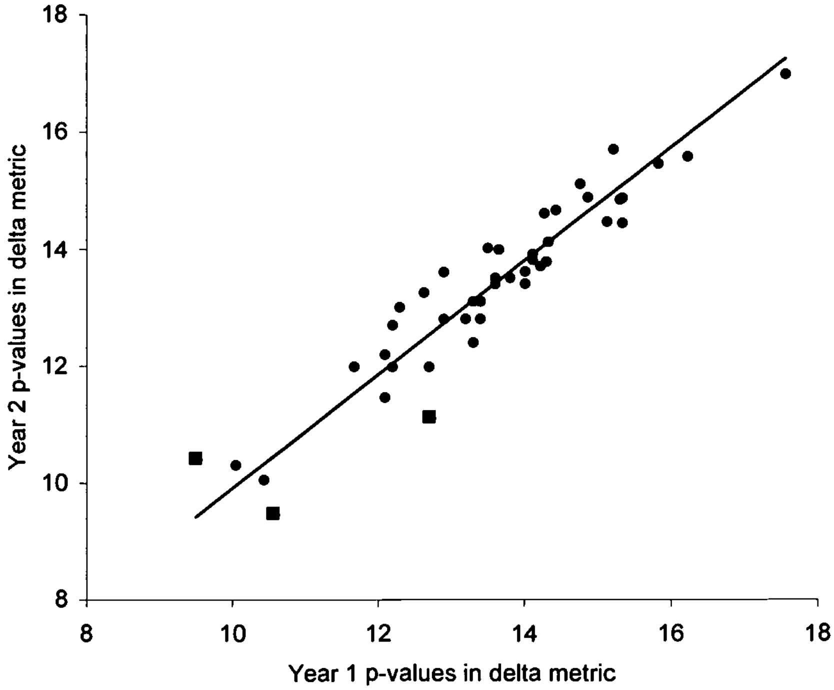

TABLE 3

Equating transformations derived under two different sets of common items

Assessment	Number of outlying common items	Transformation constants using all common items		Transformation constants using only non-outlying items in delta-plot	
		Slope	Intercept	Slope	Intercept
Mathematics 8	3	0.8479	-0.1293	0.8493	-0.1343
Science 11	3	0.7347	-0.0881	0.6887	-0.1203
Social Studies 6	2	0.9699	0.0820	0.9696	0.0844
Science 6	1	0.9035	0.0858	0.9036	0.0825

The transformation constants are derived from IRT b parameters, which as ability θ scores, are located on a scale with mean 0 and standard deviation 1. To equate the Year

2 to the Year 1 scores through equation 4, each Year 2 score was multiplied by the slope and added to the intercept.

Table 4 tabulates the average scores for the Year 1 and Year 2 assessments. The untransformed average IRT scores are not directly comparable due to IRT's scale indeterminacy. The two last columns list the equated Year 2 average scores. The difference between the equated scores obtained using all versus the delta-plot non-outlying common items is quite small. The difference is at the third decimal point[1] except in Science 11 where the difference is larger. In general, the use of the delta-plot method or not does not seem to influence greatly the equated average scores or individual scores for that matter.

TABLE 4

Untransformed and equated average test scores under a 3PL model

Assessment	Year 1 average score	Year 2 average score	Year 2 equated average score (all)	Year 2 equated average score (non-outliers)
Mathematics 8	-0.121	-0.003	-0.1319	-0.1368
Science 11	-0.126	-0.002	-0.0896	-0.1216
Social Studies 6	-0.047	-0.187	-0.0994	-0.0969
Science 6	-0.033	-0.107	-0.0109	-0.0142

Test score reporting however is not limited to average scores only. Typically, agencies report the improvement or decline of the examinee average scores to describe by how much performance has increased or decreased from one year to the next. Table 5

[1] In reality, the reporting of assessment scores entails an additional scaling step, which translates scores from the θ scale to a different one with descriptions of examinee performance attached to it. In this study, this scaling step is omitted to maintain confidentiality and in any case the reporting scale, achievement levels and cut scores differ across assessments and states. For now, suffice to look at the θ scale, which ranges from –3 to +3 with mean 0 and standard deviation 1.

tabulates the gains (or declines if negative) in average test scores under each of the two equating procedures. Gain is simply the difference of the Year 2 (equated) average score from the Year 1 average score. For presentation purposes the gains are reported in the θ scale and in a scale with a standard deviation of 100, thus the gains are multiplied by 100 and reported to the first decimal point. This second scale is similar to the well-known SAT scale, which has a mean of 500 and a standard deviation of 100. What appears in the last column is the change in gains, that is how much more or less the amount of gain in the average test scores would have been had all common items been used for equating. This ratio is calculated by dividing the gains when all common items are used by the gains when the outliers are excluded, i.e. taking the delta-plot case as the base and comparing the all-common-items case to that.

TABLE 5

Gains from Year 1 to Year 2 under the two equating item pools with a 3PL model

Assessment	Gains (all)		Gains (non-outliers)		Change in gains (ratio)
	θ scale	SD=100 scale	θ scale	SD=100 scale	
Mathematics 8	-0.0109	-1.1	-0.0158	-1.6	0.687
Science 11	0.0364	3.6	0.0044	0.4	8.363
Social Studies 6	-0.0524	-5.2	-0.0499	-5.0	1.050
Science 6	0.0221	2.2	0.0188	1.9	1.173

The difference in the amount of gains ranges from very small to quite substantial. In the case of the Social Studies grade 6 test the drop in the Year 2 average score would be similar irrespective of which of the two item clusters had been used for equating. In the case of the Science grade 11 test the reported increase would be more than eight times larger had all common items been used for performing equating; the gain would be more by 3.2 points on the scale with a standard deviation of 100.

To examine whether the type of the IRT model fitted to the data would lead to different results as regards equated average scores for Year 2 and gains from Year 1 to Year 2, the same analysis was run after item and ability parameters were obtained from a 1PL calibration. The next table corresponds to Table 4 and shows the untransformed and equated average scores after a 1PL model was fitted to the data. The difference in equated scores is in most cases to the second decimal point depending on which item cluster was used for equating.

TABLE 6

Untransformed and equated average test scores under a 1PL model

Assessment	Year 1 average score	Year 2 average score	Year 2 equated average score (all)	Year 2 equated average score (non-outliers)
Mathematics 8	0.018	0.014	0.0356	0.0349
Science 11	-0.008	0.010	0.0618	0.0209
Social Studies 6	0.010	0.006	-0.0103	-0.0107
Science 6	0.007	0.009	0.0222	0.0211

Table 7 corresponds to Table 5 and refers to the effects on gains. Again, as in the 3PL case, it seems that with a 1PL calibration the magnitude of change in gains would vary had all common items been used in equating. For Mathematics grade 8 and Social Studies grade 6, the change in gains would not be substantial. For Science grade 11 the increase in the mean score from Year 1 to Year 2 would be 2.9 points (on the scale with a standard deviation of 100) if the delta plot outlying points were discarded from the common item pool; however, if all common items had been used to generating the equating transformation that increase would be more than twice as large, 7.0 points. Overall, the change in gains is more salient under a 3PL than under a 1PL calibration.

TABLE 7

Gains from Year 1 to Year 2 under the two equating item pools with a 1PL model

Assessment	Gains (all)		Gains (non-outliers)		Change in gains (ratio)
	θ scale	SD=100 scale	θ scale	SD=100 scale	
Mathematics 8	0.0178	1.8	0.0169	1.7	1.036
Science 11	0.0698	7.0	0.0289	2.9	2.412
Social Studies 6	-0.0203	-2.0	-0.0207	-2.1	0.981
Science 6	0.0152	1.5	0.0141	1.4	1.077

Conclusions

This study explored the effects of common item selection on aggregate score results. Two clusters of common items were considered in each of four statewide assessments: use of all common items to produce the equating transformation and use of those that do not indicate large anomalous behavior on the delta plot. Excluding items on which student cohorts perform too differentially compared to their performance on other items is an issue of face validity and fairness. Given that the delta-plot procedure is just a handy method to identify the few items that behave anomalously across cohorts of students, and that the "three standard deviations away from the line" rule is just a convention, it might deserve more consideration whether the delta-plot outliers should be discarded from or kept in the common item pool.

Very few common items, less than 7% of the common item pool in all the cases presented above, were identified as "misbehaving". But excluding them from equating seemed to often have considerable impact on the average gains in proficiency from one year to the next. The effects may not be very large for individual examinee scores considering the standard errors associated with individual scores. The difference of the score of an examinee on the mean of the Year 2 distribution was similar irrespective of which common item pool was used for equating. However, in the case of a measure of annual improvement for a group (the state average) is more critical; the standard errors of group means are much smaller than those of individual scores, and thus even slight changes from one administration to the next are noteworthy. When it came to the evaluation of the gains from Year 1 to Year 2, the common item pool used to perform equating made a difference in most of the assessments analyzed. The improvement (or decline) in average state performance across years, varied considerably just by keeping in versus discarding one, two, or three outlying points of the delta plot.

This finding is significant particularly in the light of the high stakes in large-scale testing. Test scores are increasingly used for holding schools accountable. Rewards and sanctions are tied to annual progress, and therefore the gains at an aggregate level become critical. But reported gains or losses are sensitive to minor changes, such as the discarding of a few items from the common item pool. Psychometric decisions and practical techniques used in equating have had substantial impact on the magnitude of gains in this study, irrespective of which IRT model was fitted to the data. As a consequence, the accuracy of gains renders their use for assessing adequate annual progress questionable. Further research could address the development of more formal methods and more rigorous frameworks for assessing the quality of common items and identifying which items to keep or discard from the equating pool, and provide the means for making more justifiable psychometric decisions.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D.C.: AERA.

Angoff, W. H. (1972, Sept.). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069686)

Bock, R. D., Muraki, E., and Pfeiffenberger, W. (1988). Item Pool Maintenance in the Presence of Item Parameter Drift. *Journal of Educational Measurement, 25*(4), 275-285.

Dorans, N. J., and Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hambleton, R. K., Swaminathan, H., and Rogers H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications, Inc.

Kolen, M. J., and Brennan, R. L. (1995). *Test Equating Methods and Practices.* New York: Springer.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Muraki, E., and Bock, R. D. (1997). *PARSCALE.* Chicago, Il.: Scientific Software International, Inc.

Miller, A. D., and Linn, R. L. (1988). *Invariance of Item Characteristic Functions With Variations in Instructional Coverage.* Journal of Educational Measurement, 25(3), 205-219.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Iowa City, IA: Psychometric Society.

Yen, W. M. (1980). The Extent, Causes and Importance of Context Effects on Item Parameters for Two Latent Trait Models. *Journal of Educational Measurement, 17*(4), 297-311.

REPRODUCTION RELEASE

(Specific Document)

TM035106

I. DOCUMENT IDENTIFICATION:

Title: Sensitivity of IRT equating to the behavior of test equating items.

Author(s): Michalis P. Michaelides

Corporate Source:

Publication Date: April 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A □	Level 2B □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Informetion Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: MMichaelide	Printed Name/Position/Title: Michalis P. Michaelides
Organization/Address: Stanford University 485 Lasuen Mall Stanford CA 94305	Telephone: 650 497 2079 FAX: E-Mail Address: michali@stanford.edu Date: 7/7/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org