DOCUMENT RESUME

ED 478 166 TM 035 060

AUTHOR	Karkee, Thakur; Lewis, Dan M.; Barton, Karen; Haug, Carolyn
TITLE	The Effect of Including or Excluding Students with Testing Accommodations on IRT Calibrations.
PUB DATE	2003-04-00
NOTE	19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Accommodations (Disabilities); *Disabilities; *Elementary School Students; Intermediate Grades; *Item Response Theory; *Middle School Teachers; Middle Schools; Reliability; Sample Size; *Testing Accommodations
IDENTIFIERS	*Calibration

ABSTRACT

This study aimed to determine the degree to which the inclusion of accommodated students with disabilities in the calibration sample affects the characteristics of item parameters and the test results. Investigated were effects on test reliability, item fit to the applicable item response theory (IRT) model, item parameter estimates, and students' scores. Data were obtained from a statewide standards-based assessment program for reading and writing in grades 4 and 7 and in mathematics and science for grade 8. One data set comprised all tested students, including those who tested with accommodations. The other set was only those students who tested without accommodations. Differential item functioning was studied for accommodated groups. The percent of students who received accommodations varied by grade, ranging between 4.2% and 8.5% of the tested population, with higher percentages in the lower grades. The accommodation used most often was extended time followed by oral presentation. There were notable effects of including or excluding students with accommodations, and statistically significant differences between the inclusive and exclusive parameter estimates indicate that choice of calibration sample does not have a significant effect on the calibration results. Effects of these differences on the mean fit of items to the IRT model are observable but very small. Effects of differences in the item parameters on test results are less notable. Overall, results show few negative effects of calibration inclusiveness. (Contains 8 tables and 16 references.) (SLD)

The Effect of Including or Excluding Students with Testing Accommodations on IRT Calibrations

Thakur Karkee
Dan M. Lewis
Karen Barton
CTB/McGraw-Hill

Carolyn Haug
Colorado Department of Education

Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA, April 2002

The opinions expressed herein are solely those of the authors and do not necessarily represent those of CTB/McGraw-Hill or Colorado Department of Education.

Introduction

Tests are design to measure student ability under standard conditions. Any compromise in the standardized administration conditions may introduce test-irrelevant sources of variance to the test, thereby violating the validity or score comparability of the test. A key psychometric issue to consider when test scores are gathered under standard and nonstandard conditions is whether or not those scores are measuring the same abilities or constructs, regardless of condition.

Studies have indicated that test accommodations, often considered nonstandard conditions, do not change the dimensionality of the test but level the playing field (Phillips, 1994; Thurlow, Elliott, & Ysseldyke, 1998; Tindal & Fuchs, 1999). For example, accommodations, such as large print or Braille editions for visually impaired persons, or allowing extra time to dyslexic students, may be the valid accommodations and are closer approximations to the standard condition. Allowing valid accommodations produces scores for students with disabilities that measure the same attribute as standard assessments measure in non-disable individuals. That is, the assessment may inaccurately measure what these students know and are able to do without the accommodations for their disabilities.

The common accommodations for various disabilities include Braille, large print, oral presentation of directions, oral presentation of the entire test, scribe, signing, assistive communication device, and extended timing. CTB (2002) classified the accommodations into three categories for use in reporting primarily based on effect of accommodations on test scores. Students with disabilities comprise 11 percent of the total K-12 enrollment in public and non-public schools (U.S. Department of Education, 2001). Some students with disabilities are not been able to participate in assessments under standard conditions, and in the past were often

excluded from large-scale assessment programs (Allington & McGill-Franzen, 1992; Ingles, 1991; McGrew, Thurlow, & Spiegel, 1993; Olson & Goldstein, 1997). Furthermore, these students were often excluded from the sample used to calibrate and scale test items primarily because the accommodations have been thought to introduce a trait-irrelevant source of variance (Phillips, 1994). Federal legislation currently requires the inclusion of all students that can meaningfully participate in large-scale testing programs including those whose IEP or 504 plans specify the use of testing accommodations.

The effect of testing with accommodations, however, is not fully understood; research has returned mixed results (Perlman et. al., 1996; McDonnell et. al, 1997). For example, NAEP (1999) reported that although the score gain reduced when the inclusion rate was increased between 1994 and 1998, the remaining gain was still statistically significant in the reading scores in Kentucky. In a study of experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities, Elliott et. al (1999) noted that when individualized testing accommodations are provided to students with disabilities, the performance of students with and without disabilities are very similar on complex math and science performance tasks. However, very little research has been conducted to determine how including or excluding students who test with accommodations in the calibration sample affects the resulting item parameter and student ability estimates.

The present study is intended to determine the degree to which the inclusion of accommodated students with disabilities in the calibration sample affects the characteristics of the item parameters and the test results. Specifically, an investigation of the effect(s) including or excluding students tested with accommodations in the calibration has upon (a) test reliability, (b) item fit to the applicable Item Response Theory (IRT) model, (c) item parameter estimates,

and (d) students' scores. In addition, differential item functioning (DIF) analyses were conducted to examine potential sources of item bias.

Data Source and Method

Data for this study were obtained from a statewide standards-based assessment program for Reading and Writing in grades 4 and 7, and in grade 8, Mathematics and Science. Two data sets were created from the tested population. The first data set is comprised of all tested students, including those who tested with accommodations referred to as the "inclusive sample," denoted N_i. The second data set is comprised of students who tested without accommodations, referred to as the "exclusive sample," and denoted N_e. Item parameter estimates were obtained, as described below, using each sample. The parameter estimates resulting from the calibration sample N_i are referred to as P_i and those resulting from N_e are referred to as P_e. Thus, four sets of analyses were conducted:

N_iP_i refers to analyses conducted using the inclusive parameters with the analyses results reported for the inclusive sample.

N_eP_e refers to analyses conducted using the exclusive parameters with the analyses results reported for the exclusive sample.

N_iP_e refers to analyses conducted using the exclusive parameters with the analyses results reported for the inclusive sample.

N_eP_i refers to analyses conducted using the inclusive parameters with the analyses results reported for the exclusive sample.

The computer program PARDUX (Burket, 1991) was used for item calibration. PARDUX simultaneously estimates parameters for multiple-choice (MC) and constructed-response (CR) items using two-parameter partial credit and three-parameter logistic IRT models

for CR and MC items, respectively. In order to facilitate comparison, a set of anchor items was used to link both inclusive and exclusive data sets into a common scale using the procedure suggested by Stocking and Lord (1983).

Model Fit

PARDUX computes Goodness-of-fit statistics for each item to examine how closely the item's data conform to the item response models. A procedure described by Yen (1981) was used to measure fit. In this procedure, students are rank ordered on the basis of their $\hat{\theta}$ values and sorted into ten cells with ten percent of the sample in each cell. Each item j in each decile i has a response from N_{ij} examinees. The fitted IRT models are used to calculate an expected proportion E_{ijk} of examinees who respond to item j in category k. The observed proportion O_{ijk} is also tabulated for each decile, and the approximate chi-square statistic is given by

$$Q_{1j} = \sum_{i=1}^{10} \sum_{k=1}^{m_j} \frac{N_{ij}(O_{ijk} - E_{ijk})^2}{E_{ijk}},$$

Q_{1j} should be approximately distributed as a chi-square with degrees of freedom (DF) equal to the number of "independent" cells, $10(m_j\text{-}1)$, minus the number of estimated parameters. For the 3PL model m_j =2, so $DF = 10(2-1) - 3 = 7$. For the 2PPC model, $DF = 10(m_j - 1) - m_j = 9m_j - 1$. Since DF differs between MC and PA items and between PA items with different score levels m_j, Q_{1j} is transformed, yielding the test statistic

$$Z_j = \frac{Q_{1j} - DF}{\sqrt{2DF}}.$$

This statistic is useful for flagging items that fit relatively poorly. Z_j is sensitive to sample size, and critical values for flagging an item based on Z_j have been developed. The

critical values for N_i and N_e are included with the detailed results in Table 3, which contains the mean Z values for the items in each grade/content area computed for each sample/parameter combination. Higher Z values indicate poorer fit to the IRT model and values beyond the critical Z indicate poor fit.

Differential Item Functioning (DIF)

The Linn and Harnisch (1981) procedure, generalized to include constructed response items, was used to study DIF for accommodated groups. The parameters for each item (a_i, b_i, and c_i) and the trait or scale score (θ) for each examinee are estimated for the three-parameter logistic model:

$$P_{ij}(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta_j - b_i)]},$$

where $P_{ij}(\theta)$ is the probability that examinee j, with a given value of θ, will obtain a correct score on item i. Note that the item parameter estimates are based on data from the total sample of valid examinees. The sample is then divided into groups, and the members in each group are sorted into ten equal score categories (deciles) based upon their location on the score scale (θ). The expected proportion correct for each group based on the model prediction is compared to the observed (actual) proportion correct obtained by the group. The proportion of people in decile g who are expected to answer item i correctly is

$$P_{ig}(\theta) = \frac{1}{n_g} \sum_{j \varepsilon g} P_{ij}(\theta),$$

where n_g is the number of examinees in decile g. The proportion of students expected to answer item i correctly (over all deciles) for a group is given by:

$$P_{i.} = P_i(\theta). = \frac{\sum_{g=1}^{10} n_g P_{ig}(\theta)}{\sum_{g=1}^{10} n_g}.$$

The corresponding observed proportion correct for examinees in a decile (O_{ig}) is the number of examinees in decile g who answered item i correctly divided by the number of people in the decile (n_g). That is,

$$O_{ig} = \frac{\sum_{j \varepsilon g} u_{ij}}{n_g},$$

where u_{ij} is the dichotomous score for item i for examinee j.

The computation of the observed proportion answering each item correctly (over all deciles) for a complete group is given by the formula:

$$O_{i.} = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g}.$$

After the values are calculated for these variables, the difference between the observed proportion correct and expected proportion correct can be computed. The decile group difference (D_{ig}) for observed and expected proportion correctly answering item i in decile g is

$$D_{ig} = O_{ig} - P_{ig},$$

and the overall group difference (D_i) between observed and expected proportion correct for item i in the complete group (over all deciles) is

$$D_{i.} = O_{i.} - P_{i.}.$$

These indices are indicators of the degree to which group members perform better or worse than expected on each item, based on the parameter estimates from all sub-samples.

Results

The percent of students who received testing accommodations varied by grade, ranging from 4.2 to 8.5 percent of the tested population with higher percentages in lower grades. The most common accommodation given was generally Extended Time followed by Oral Presentation. For grade 8 science, the percentage of students receiving Oral Presentation (2.4%) is slightly higher than those receiving Extended Time (1.5%). See Table 1 for details of the accommodation rates.

Model Fit

A goodness-of-fit statistic (Z), described by Yen (1981), was computed for each item to examine how closely the responses fit the associated IRT model. The critical Z values for N_i and N_e in each grade/content area are included with the detailed results in Table 2, which contains the mean Z values for the items in each grade/content area. Poor fit is indicated when an item's Z value exceeds the critical Z.

The mean fit-statistic, being sensitive to sample size, was compared only for the same samples. For the inclusive sample, the results were non-uniform. For two of the six grade/content areas (Grade 4 Reading and Grade 8 Science), the mean fit was better using the inclusive parameters. For Grade 4 Writing, Grade 7 Reading, and Grade 8 Math, the mean fit was worse; and for Grade 7 Writing, the mean fit was the same as when using inclusive parameters. However, the result for the exclusive sample was uniform. The mean fit statistics were slightly larger across all grade/content areas when using the inclusive parameters, showing a poorer fit, than using the exclusive parameters. Note that the mean fit values were substantially lower than the critical values in all cases across both inclusive and exclusive parameters.

Individual item fit results are summarized in Table 3. The same items exhibited less than optimal fit in each grade/content area irrespective of inclusive or exclusive population and

parameters except for grade 8 Mathematics. Increased fit values for the inclusive and exclusive samples were non-uniform. For the inclusive sample, four items (one item in Grade 4 Reading, two items in Grade 8 Mathematics, and one item in Grade 8 Science) had smaller fit values for poorly fitted items due to the inclusive parameters compared to those same four items due to the exclusive parameters. Similarly, for the exclusive sample, two of the items at Grade 8 (one in Mathematics and one in Science) had smaller fit values for poorly fitted items due to the inclusive parameters compared to those same two items due to the exclusive parameters. Note that most of the poorly fitted items were constructed-response items.

Item Parameters

The item calibration software PARDUX (Burket, 1991) provides estimates of the error associated with the parameter estimates. The error estimates were used to test whether the differences between the parameters estimated by N_iP_i and N_eP_e were statistically significant at the .05 level. Ninety-five percent confidence intervals were constructed about each of the parameter estimates for P_i to determine whether the corresponding parameter for P_e was contained in that interval. Table 4 shows the percent of items whose "a" (discrimination), "b" (location), or "c" (pseudo-guessing) parameters were different at the .05 level of significance. An average of 43%, 66%, and 9% of the a, b, and c parameters, respectively, showed statistically significant differences across the grade/content areas.

Ability Estimates

Both sets of parameters in the logit metric were placed on a common scale through a set of anchor items using a procedure developed by Stocking and Lord (1983). Ability estimates were computed for the two samples, N_i and N_e, by scoring the students' responses with P_i and P_e. The mean scale scores for each sample remained quite stable when scored under either parameter

set (Table 5). The differences in mean scale scores for a given sample (N_i or N_e) scored under each parameter set (P_i or P_e) ranged in absolute value from 0.06 to 0.66.

The impact of inclusion or exclusion of the accommodated students in the calibration sample to the classification of students in different percentiles or proficiency levels is displayed in Table 6. It shows reasonably stable scale scores for Reading and Writing at Grades 4 and 7 across percentiles with a maximum difference of 1 scale score point and virtually no differences at the 50^{th} percentile. The scale score difference in percentiles is slightly higher for Mathematics and Science at Grade 8 ranging from 0 to 5 scale score points with a difference of 3 scale score points at the 50^{th} percentile.

Traditionally, item parameters from exclusion sample (P_e) has been used to score responses for population (N_i) and generate corresponding percentiles scale scores for reporting. Therefore, only the percentiles at proficiency level cut-points for N_iP_e is compared with N_iP_i here (Table 7). Results indicate that the percentile at the cut-points is same at all proficiency levels for the Reading and Writing and at proficiency levels 3 and 4 for the Mathematics and Science for both populations. The percentile at proficiency levels 1 and 2 for the Mathematics and Science decreased for N_iP_i. The difference, however, was small ranging from 1 to 2 percentile points.

Differential Item Functioning (DIF)

DIF was assessed using a modification of the procedure suggested by Linn and Harnisch (1981) in which differences between the observed and expected results were compared for focal group students in each decile. Small sample sizes limited the DIF analyses to three focal groups. These focal groups consisted of students provided the following accommodations: (a) oral presentation of test, (b) extended time, and (c) any accommodation. Analyses were conducted by comparing the numbers of items indicated as manifesting DIF under each set of parameters

(P_e and P_i). The results appear in Table 8. In general the direction of the DIF was not uniformly in favor of or against the focal groups, however more DIF items favored the focal groups.

For students having the accommodation "oral presentation," the use of parameters P_i tended to reduce the number of DIF items, as would be expected. Similar results are observed for students receiving "extended time." For students having "any accommodation," no items were indicated as manifesting DIF for either parameter set for grades 4 and 7 Reading and Writing and grade 8 Mathematics. In grade 8 Science, there was one DIF item under each parameter set.

Summary and Discussion

The most referred accommodation used in this state's testing program was extended time followed by the oral presentation, which is congruent to the previous findings that extended time is the frequently offered accommodation type (NAEP, 2000). There are notable effects of including or excluding students who test with accommodations in calibration samples. Statistically significant differences between the inclusive and exclusive parameter estimates indicate that the choice of calibration sample does have a significant effect on the calibration results.

The effects of these differences on the mean fit of the items to the IRT model are observable, but very small. At the item level, one item was flagged as exhibiting poor fit under the inclusive parameters that did not show poor fit under the exclusive parameters. Although the fit-statistic for the inclusive sample was increased, the critical Z value was larger for the inclusive sample as it is a function of sample size.

The effects of the differences in the item parameters on the test results, however, are less notable. Of concern to most test users is the effect on test scores. Mean test scores were highly

stable when students were scored with either inclusive or exclusive parameters, varying by less than one scale score point for either the inclusive or exclusive samples. Also, the stable scale scores at different percentiles and percentage of students in different proficiency levels suggest that the impact of including accommodated students in the calibration sample was visible but minimal on student test scores. The results also implies that the mis-classification of students in different proficiency levels due to NiPi is minimal in all grades and content areas with more visible in the Mathematics and Science.

A test can not be valid with many items that exhibit bias, or DIF, for or against a given focal group. Thus, it is notable, but not unexpected, that the use of the inclusive parameters tends to decrease the number of items exhibiting DIF for or against students who test with extended time or the use of a scribe. However, for the focal group of students who tested with any accommodation, only one item exhibited DIF across all six grade/content areas. This was unchanged regardless of inclusive or exclusive parameters used.

Under the current, inclusionary, testing practices mandated by federal legislation, one would assume that these students would also be included in IRT calibration samples, unless cause can be shown to exclude them. The results of the present study showed few negative effects of calibration inclusiveness. Further research is warranted to replicate and extend the above results, but given the results of the current study, the authors would support the inclusionary practice in item calibrations.

Additionally, the results are based on large sample sizes. The impact of accommodated students on the test score may vary if the proportion of the accommodated students and sample sizes varied. A simulation study with systematic variation of percentage of accommodated

students in the sample as well as variation of sample sizes would be appropriate for future research.

References

Allington, R., L., & McGill-Franzen, A. (1992). Unintended effects of educational reform in New York. *Educational Policy, (4),* 397-414.

Burket, G. R. (1991). PARDUX [Computer program]. Unpublished.

CTB (2002). Guidelines for inclusive test administration. CTB/McGraw-Hill. Unpublished.

Ingles, S. J. (1991, April). *The problem of excluded baseline students in a school-based longitudinal study: Correcting national dropout estimates and accommodating eligibility change over time.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Linn, R. L. & Harnisch, D. (1981). Interaction between item content and group membership in achievement test items. *Journal of Educational Measurement*, 18, 109-118.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington, DC: National Research Council.

McGrew, K. S., Thurlow, M. L., & Spiegel, A. N. (1993). An investigation of the exclusion of students with disabilities in national data collection programs. *Educational Evaluation and Policy Analysis, 15*, 339-352.

NAEP (2000). Increasing the Participation of Special Needs Students in NAEP. A Report on 1996 NAEP Research Activities.

NAEP (1999). Increasing the Participation of Special Needs Students in NAEP. National Report Card 1994-1998.

Olson, J., & Goldstein, A. (July 1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress.* National Center for Education Statistics, U.S. Department of Education, Office of Educational Research and Improvement.

Perlman, C., Borger, J., Collins, C., Elenbogen, J., & Wood, J. (1996). *The effect of extended time limits on learning disabled students' scores on standardized reading tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, April 9, 1996.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, 7(2), 94-120.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Tindal, G. & Fuchs, L. (1999). A summary of research of test accommodations: What we know so far. Mid-South Regional Resource Center, University of Kentucky.

U. S. Department of Education (2001). Percentage (Based on Estimated Resident Population) of Children Ages 6-17 Served Under IDEA, Part B By Disability During the 1999-2000.

Yen, W. M (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 24, 185-201.

Table 1. The percent of students receiving various accommodations

Accommodation	Grade 4		Grade 7		Grade 8	
	Reading	Writing	Reading	Writing	Mathematics	Science
Sample size (N)	44,403	43,725	35,977	35,254	52,225	52,140
No accommodation or missing data	91.5	92.6	94.8	95.1	95.0	95.8
Large print	0.0	0.0	0.1	0.0	0.1	0.0
Braille version	0.0	0.0	0.0	0.0	0.1	0.1
Oral presentation	1.5	1.6	1.9	1.9	2.3	2.4
Use of number line	0.1	0.1	0.0	0.0	0.0	0.0
Scribe	0.4	0.5	0.2	0.3	0.2	0.2
Signing	0.0	0.0	0.0	0.0	0.0	0.0
Assistive communication device	0.0	0.0	0.0	0.0	0.0	0.0
Extended time	6.4	5.1	3.0	2.7	2.3	1.5
Oral presentation of Entire Test	0.0	0.0		0.0		0.0
Total Receiving Any Accommodation:	8.5	7.4	5.2	4.9	5.0	4.2

Table 2. Mean Model Fit-statistics (Z)

Grade/ Content Area	Inclusive Sample			Exclusive Parameters		
	Mean value of Z for N_iP_i	Critical Z for N_i	Mean value of Z for N_iP_e	Mean value of Z for N_eP_e	Critical Z for N_e	Mean value of Z for N_eP_i
G4 Reading (70)*	16.25 (44,403)**	118.41	17.15 (44,403)	13.81 (40,596)	108.26	14.60 (40,596)
G4 Writing (44)	17.10 (43,725)	116.60	16.99 (43,725)	14.52 (40,496)	107.99	15.57 (40,496)
G7 Reading (77)	11.22 (35,977)	95.94	10.90 (35,977)	9.30 (34,108)	90.95	10.24 (34,108)
G7 Writing (59)	18.32 (35,254)	94.01	18.32 (35,254)	16.33 (33,527)	89.41	17.33 (33,527)
G8 Math (60)	26.57 (52,225)	139.27	26.32 (52,225)	22.55 (49624)	132.33	23.57 (49,624)
G8 Science (73)	15.68 (52,140)	139.04	15.71 (52,140)	13.75 (49,955)	133.21	14.29 (49955)

* Total number of items

** Sample size

Table 3. Model Fit-statistics (Z) at Item Levels

Grade/ Content Area	Inclusive Sample (N_i)			Exclusive Sample (N_e)		
	P_i	Critical Z for N_i	P_e	P_i	Critical Z for N_e	P_e
G4 Reading (70)	I-64* (138.3)	118.4 (44,399)	I-64* (167.7)	I-64* (139.9)	108.2 (40,592)	I-64* (138.3)
G4 Writing (44)	I-8 (235.2)	116.4 (43,666)	I-8 (223.3)	I-8 (206.2)	107.8 (40,438)	I-8 (190.6)
G7 Reading (77)	∅	95.9 (35,977)	∅	∅	91.0 (34,108)	∅
G7 Writing (59)	I-33 (114.1) I-59 (409.5)	94.0 (35,250)	I-33 (108.1) I-59 (399.3)	I-33 (105.4) I-59 (389.5)	89.4 (33,523)	I-33 (92.5) I-59 (367.2)
G8 Math (60)	I-29* (191.5) I-44 (151.2)	139.3 (52,221)	I-29* (195.9) I-44* (154.4)	I-29* (141.1)	132.3 (49,620)	I-29* (142.3)
G8 Science (73)	I-61 (186.8)	139.0 (52,140)	I-61 (194.9)	I-61(146.7)	133.2 (49955)	I-61 (156.1)

*Multiple-Choice Item,
I stands for item. Values in parentheses under Grade are total number of items; under P_i and P_e are Z-statistics; and under Critical Z for N_i or N_e are sample size.

Table 4. Number and Percent of parameters estimated under P_i that are different from those estimated under P_e at the .05 level of significance

Grade/ Content Area	Total Items (MC items)	Discrimination (a) Number (percent) different at the .05 level	Difficulty (b) Number (percent) different at the .05 level	Pseudo-guessing (c) Number (percent) different at the .05 level
G4 Reading	70 (53)	44 (63%)	48 (91%)	10 (19%)
G4 Writing	44 (30)	18 (41%)	26 (87%)	5 (17%)
G7 Reading	77 (60)	28 (36%)	36 (60%)	4 (7%)
G7 Writing	59 (42)	35 (59%)	30 (71%)	2 (5%)
G8 Math	60 (45)	20 (33%)	24 (53%)	1 (2%)
G8 Science	73 (58)	19 (26%)	20 (34%)	1 (2%)
Total (Average %)		164 (43%)	184 (66%)	23 (9%)

*Difficulty and Pseudo-guessing differences were estimated for MC items only.

Table 5. Ability Estimates (Standard Deviation)

Grade/ Content Area	N_iP_i	N_iP_e	N_eP_e	N_eP_i
G4 RD	256.47 (45.16)	256.99 (44.74)	260.13 (43.17)	259.75 (43.11)
G4 WR	259.40 (42.76)	259.18 (42.85)	261.79 (41.73)	261.71 (41.61)
G7 RD	249.98 (50.59)	250.09 (49.68)	253.23 (47.44)	253.73 (47.78)
G7 WR	248.34 (49.37)	248.02 (48.54)	251.03 (46.82)	251.09 (47.55)
G8 MA	245.21 (57.26)	245.00 (57.50)	248.00 (55.50)	248.31 (55.20)
G8 SC	247.16 (55.99)	246.50 (56.00)	249.50 (54.00)	249.75 (53.92)

Note: Ability estimates are on a scale score metric (Mean = 250, SD = 50).

Table 6. Ability Estimates by Percentiles

Percentile	Grade 4						Grade 7					
	Reading			Writing			Reading			Writing		
	N_iP_e	N_iP_i	Diff.	N_iP_e	N_iP_i	Diff.	N_iP_e	N_iP_i	Diff.	N_iP_e	N_iP_i	Diff.
10	199	199	0	206	206	0	186	185	-1	186	185	-1
25	231	230	-1	233	233	0	220	220	0	217	216	-1
50	260	260	0	259	259	0	254	254	0	249	249	0
75	287	287	0	286	286	0	284	285	1	281	281	0
90	309	309	0	312	312	0	310	311	1	308	309	1

Percentile	Grade 8					
	Mathematics			Science		
	N_iP_e	N_iP_i	Diff.	N_iP_e	N_iP_i	Diff.
10	176	176	0	174	179	5
25	212	217	5	214	217	3
50	251	254	3	251	254	3
75	284	286	2	285	286	1
90	313	314	1	313	314	1

Table 7. Percentiles at Proficiency Level cut-points for the Inclusive Sample

Content	Grade	Proficiency Levels*	Cut-Points	N_iP_e Percentiles	N_iP_i Percentiles
Reading	4	1	444 and below	8-9	8-9
		2	445 – 495	36	36
		3	496 – 561	90-91	91
		4	562 and above	91<	91<
Writing	4	1	468 and below	15-16	15-16
		2	469 – 521	62	61
		3	522 – 587	96-97	96-97
		4	588 and above	97<	97<
Reading	7	1	445 and below	13	13
		2	446 – 493	41	41
		3	494 – 580	96-97	96
		4	581 and above	97<	96<
Writing	7	1	399 and below	2-3	2-3
		2	400 – 508	58	57-58
		3	509 – 633	99<	99<
		4	634 and above	99<	99<
Math	8	1	474 and below	31-32	29
		2	475 – 521	65-66	64
		3	522 – 562	89-90	89-90
		4	563 and above	90<	90<
Science	8	1	455 and below	20-21	19
		2	456 – 505	53	51
		3	506 – 581	95-96	95-96
		4	582 and above	96<	96<

*1=Unsatisfactory, 2=Partial Proficient, 3=Proficient, 4=Advanced

Table 8. Number of DIF Items in the Inclusive and Exclusive samples using the inclusive and exclusive parameters.

Grade	Focal Groups: Oral Presentation		Extended Time		Any Accommodation	
	N_iP_e	N_iP_i	N_iP_e	N_iP_i	N_iP_e	N_iP_i
G4 Reading	2+*	1+	1+	1+		
G4 Writing	2-	1-				
G7 Reading	1+	1+	1+	1+		
	1-			1-		
G7 Writing	2+	1+	2+	1+		
G8 Math	4+	1+	1+			
G8 Science	3-	3-	2-	2-	1-	1-

* A + (-) after the number of items indicates that the DIF was manifest in favor of (against) the focal group.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM035060

I. DOCUMENT IDENTIFICATION:

Title: The Effect of Including or Excluding Students with Testing Accommodations on IRT Calibration.

Author(s): ~~Karkee~~, Thakur Karkee, Dan M. Lewis, Karen Barton, & Carolyn Haug

Corporate Source: CTB/McGraw-Hill
Colorado Department of Education

Publication Date: Paper Presented in 2002 NCME

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

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

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

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

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 Information 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: Thakur Karkee	Printed Name/Position/Title: Thakur Karkee / Research Scientist
Organization/Address: CTB/McGraw-Hill Companies LLC. ~~Colorado~~ 20 Ryan Ranch Rd. Monterey, CA 93940	Telephone: (831) 393-7317 FAX: (831) 393-7016
	E-Mail Address: tkarkee@ctb.c Date: 6/12/03

(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