EDUCATIONAL ASSESSMENT, *10*(4), 357–375

Improving Assessment Validity for Students With Disabilities in Large-Scale Assessment Programs

Julia Shaftel
Center for Psychoeducational Services
University of Kansas

Xiangdong Yang, Douglas Glasnapp, and John Poggio
Center for Educational Testing and Evaluation
University of Kansas

A test designed with built-in modifications and covering the same grade-level mathematics content provided more precise measurement of mathematics achievement for lower performing students with disabilities. Fourth-grade students with disabilities took a test based on modified state curricular standards for their mandated statewide mathematics assessment. To link the modified test with the general test, a block of items was administered to students with and without disabilities who took the general mathematics assessment. Item difficulty and student mathematics ability parameters were estimated using item response theory (IRT) methodology. Results support the conclusion that a modified test, based on the same curricular objectives but providing a more targeted measurement of expected outcomes for lower achieving students, could be developed for this special population.

The 1997 amendments to the Individuals with Disabilities Education Act (IDEA; 1997) and the No Child Left Behind Act of 2001 (NCLB; 2002) require that all students participate in statewide accountability assessments, including all students with disabilities. Under both federal programs the majority of students with disabilities is expected to participate in general assessments because this promotes greater instructional opportunity and higher achievement expectations for students who have historically been exempted from accountability testing and hence from

Correspondence should be addressed to Julia Shaftel, Center for Psychoeducational Services, University of Kansas, 1122 West Campus Road, Room 130, Lawrence, KS 66045. E-mail: jshaftel@ku.edu

measurement of their learning (Thurlow, Elliott, & Ysseldyke, 1998). However, general assessments may not be as valid for all students with disabilities due to the lack of correspondence between appropriate instruction and items on the assessment. Therefore, in addition to general curricular assessments, alternative forms of tests, as well as a variety of accommodations and modifications, have been proposed and developed.

ACCOMMODATIONS VERSUS MODIFICATIONS

Accommodations have been defined as those alterations to test presentation, setting, timing, scheduling, and response that mitigate the barrier of disability and allow a student with disabilities to demonstrate actual achievement in a particular academic area without changing the underlying construct of what is being measured (Hollenbeck, Tindal, & Almond, 1998; Schulte, Elliott, & Kratochwill, 2001; Thurlow et al., 1998). Appropriate accommodations are those that are used regularly for instruction and tailored to individual learning needs. Ideally, accommodations selectively benefit students with special needs without conferring an undue advantage; students without those needs would not experience a benefit from the accommodation (Hollenbeck et al., 1998; Schulte et al., 2001). For this reason, some alterations used frequently for instruction, such as oral presentation of reading passages or using a calculator for computation items, may not be permitted during testing without calling into question the meaning of the construct measured by the assessment. Due to the controversial nature of some accommodations and their unknown impact on test score comparability, the selection and use of accommodations for special needs populations is currently the topic of a great deal of research (Destefano, Shriner, & Lloyd, 2001; Johnson, Kimball, Brown, & Anderson, 2001; Johnson & Monroe, 2004; Schulte et al., 2001; Tindal & Fuchs, 2000). Alterations that are likely to change the nature of what is being tested have been called modifications to distinguish them from accommodations that are believed to preserve score comparability (Hollenbeck et al., 1998; Schulte et al., 2001).

In Kansas, the terms *accommodation* and *modification* are not used interchangeably. A clearly defined set of accommodations, such as extra time, frequent breaks, oral presentation of nonreading comprehension items, and dictation of answers, is available to any student depending on individual need and regular instructional use, not on disability status or label. Modifications that may change the nature of the test and limit score comparability are not permitted except in certain circumstances. For example, calculator use on mathematics assessments is not permitted at fourth grade except for students with disabilities who have that modification noted in their Individualized Education Plans (IEPs). Some modifications, such as oral presentation of reading passages, are prohibited for all students.

Even multiple accommodations may not provide sufficient access to the general instructional curriculum or corresponding assessments for some lower performing students with disabilities. This is particularly true where accountability assessments are designed to be rigorous and represent coverage over a grade-appropriate range of curricular content. Yet current legislation, including the IDEA 1997 amendments and NCLB, asserts the right of students with disabilities, except for a very small minority of students with the most significant challenges, to obtain access to the general curriculum and to be assessed on that curriculum. When the general assessment is too difficult, how can these students be assessed in a manner that respects their achievement and provides valid and interpretable scores?

During the 2000–2001 academic year, all students, including those with disabilities, participated in our state's accountability assessments. For students with the most significant disabilities an Alternate Assessment was offered, consisting of a portfolio demonstrating performance during the year and a rating scale completed by interview in the spring. The eligibility criteria for the Alternate Assessment were intentionally quite restrictive: A student must be instructed in a curriculum that corresponds to the state's Extended Curricular Standards in Reading, Writing and Mathematics, which are downward extensions of the state's general curricular standards in these subject matter areas. Students participating in the Alternate Assessment during the first year of implementation comprised only about 0.75% of all students assessed.

Because the constraints on instructional curriculum and eligibility for the Alternate Assessment resulted in such a small proportion of eligible students, a pool of students with disabilities remained for whom the general assessment was still too difficult, did not correspond well with their adapted curricular needs, and hence lacked validity as a measure of their skills. In mathematics, our state chose to develop a third option for this "gray area" of students: a series of assessments based on general curricular standards but with specific content and performance modifications built in. The purpose of these new assessments was not just to develop easier tests but to develop assessment instruments suited to the curriculum and instruction of eligible students with disabilities so that their mathematics achievement could be validly measured rather than simply falling at the floor of the distribution on the general test. One of these assessments was the modified mathematics assessment designed for lower performing fourth-grade students with disabilities who were not eligible for the portfolio/rating scale Alternate Assessment. For this project, that modified assessment was selected for more intensive study and comparison with the general assessment.

A major goal of the modified mathematics assessment was to retain grade-level-specific indicators for evaluation, thus matching the content of the modified test with that of the general assessment. A second major purpose was to control and regularize the types of modifications that students with disabilities may need by building them into the test, thereby ensuring standardized presenta-

tion and score comparability. Including the modifications within the standardized test protocol should minimize the validity problems of allowing modifications to be developed locally by IEP teams, which may then result in assessments given in myriad nonstandardized ways for which scores cannot be meaningfully compared. This research evaluated the construction and use of the fourth-grade modified mathematics test with the following general questions in mind: Can an assessment with built-in curricular modifications be an effective tool for evaluating the achievement of lower performing students with disabilities in the general grade-level curriculum? How do scores on the modified assessment compare to scores on the general assessment instrument? Who should take the modified test instead of the general assessment?

METHOD

Modified Assessment Instrument

Instructional objectives, referred to as indicators, were defined by the state for students with disabilities. These included all of the conceptual content of the general curricular indicators. The general curricular indicators for mathematics at each grade had been previously approved by the state school board and reviewed by an impartial agency, the Thomas B. Fordham Foundation, which found them to be comprehensive and of high quality (Finn & Petrilli, 2000). Starting from this basis, teams of experienced mathematics and special education teachers working under the direction of the state's department of education met to review the indicators and determine what, if any, changes were necessary to meet the curricular and learning needs of students with disabilities. The team reviewed each indicator measured on the state large-scale mathematics assessments for appropriateness for the modified tests. Changes were made to existing mathematics curricular indicators at each tested grade level (4th, 7th, and 10th grades) in a variety of ways, including

1. Simplifying operands, such as restricting computation to whole numbers instead of decimals or limiting the number of decimal places to be manipulated.
2. Limiting the number of steps or operations to be performed.
3. Limiting abstract content by requiring that items be relevant to students with disabilities.

In addition to indicator modifications, test and item alterations for the modified tests were also defined by the state's department of education in conjunction with the teacher teams:

4. Reducing the total number of items on the test.
5. Removing extraneous information from word problems.

6. Simplifying the language or wording of the problem.
7. Including key definitions and formulas within the problem.
8. Permitting calculators to be used throughout the assessment.

Some of these last alterations may be more accurately categorized as accommodations based on recent research. Simplified language on math items may assist low-achieving students or those with disabilities but does not consistently improve the scores of general education students and English Language Learners (Abedi, Lord, & Plummer, 1997; Johnson & Monroe, 2004; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2002, 2003). Calculator use has not been shown to have an overall beneficial effect on math scores for any group of students, with the impact of calculators limited to certain types of items (Shaftel et al., 2002, 2003; Tindal & Fuchs, 2000). Nonetheless, these alterations were considered to be possible test modifications and were treated as such for the purpose of the development of the assessment described here.

Examples of original and modified indicators for all four mathematics standards—Numbers and Computation, Algebra, Geometry, Data—are shown in Table 1. Comparison of these examples shows that the numerical, computational, and algebraic concepts included in the content to be assessed were not changed but the application of that content to test problems was limited and simplified. This close correspondence between the original and modified indicators is evidence of the alignment between the modified test and the general assessment, a key issue for assessing students with disabilities on general curricular content. The modified indicators then provided guidance to teacher item-writing teams on the types of problems and content that could be included in the modified tests. Alignment with the general standards meant that many existing test items, most with minor changes, were appropriate for use on the modified test, as described later.

After the indicator, test, and item modifications had been defined by the state's department of education teacher teams, a pool of test items was prepared and reviewed for alignment with the standards, this time by qualified teacher teams under the direction of the test contractor. Most of the test items were selected from the general education assessment item pool on the basis of good statistical item properties, including high point-biserial correlations and high p values, typically .6 and above, which indicated that they were effective for lower performing students. After revisions, the pool of modified items was field-tested with students with disabilities before final selection of items for the modified tests was made.

The modified assessment at the fourth-grade level consisted of 35 items rather than the 52 items in the general assessment. The table of specifications defining assessment coverage of the domain was the same as for the general assessment, ensuring the same relative content emphasis but with fewer items to assess each standard. Of the 35 items prepared for the fourth-grade test, 22 items were mathematically identical to items in the general assessment but with simplified wording and, in some cases, simplification of accompanying diagrams or illustrations. In

TABLE 1
Original and Modified Mathematics Indicators

Content	*Original Indicator*	*Modified Indicator*
Number and computation: number sense	Determines reasonableness of numerical values involving whole numbers to 1,000,000, simple fractions, and decimals to the thousandths.	Determines reasonableness of numerical values involving whole numbers to 1,000,000, simple fractions, and decimals to the hundredths.
Number and computation: computation	Performs whole number division using dividends with up to three digits and a one-digit divisor.	Performs whole number division without remainders using dividends with up to three digits and a one-digit divisor.
Algebra: variables, equations, and inequalities	Formulates and solves problem situations involving one-step equations in one unknown with a whole number solution.	Solves one-step equations involving one unknown with a whole number solution such as finding any missing number in a multiplication or division equation based on the multiplication and division facts for numbers up to 12 × 12, equations involving money such as 8 quarters + 10 dimes = Δ dollars and 100 × Δ = 600. Geometric figures such as a square or triangle will be used to represent the unknown.
Geometry: measurement and estimation	Formulates and solves real-world problems by applying measurements and measurement formulas.	Solves real-world problems by applying measurements and measurement formulas.
Data: statistics	Uses data analysis to make reasonable inferences, decisions, predictions, and to develop convincing arguments from data displayed in a variety of formats: frequency tables horizontal and vertical bar graphs Venn diagrams or other pictorial displays charts and tables line graphs pictographs	Uses data analysis to make accurate inferences from data displayed in a variety of formats: frequency tables horizontal and vertical bar graphs Venn diagrams for up to two attributes charts pictographs

addition to language modifications, five additional items were mathematically altered from items in the original item pool as described in the modified mathematics indicators by simplifying place values, interpretation of Venn diagrams, and computations involving time and money. Three entirely new items were prepared because existing items written for specific instructional indicators could not be used to assess the corresponding modified indicators. The remaining five items consisted of three drawn unchanged from the general assessment and two items originally prepared for the general assessment but not used on any of the test forms.

In addition to the specific modifications listed previously, several presentation and response alterations best classified as accommodations were available to all students who took the modified assessment. Modified test booklets had fewer items per page, increased font size, and some additional illustrations that did not supply information needed to solve a problem. Students were tested in small groups by their special education teachers rather than as part of a general mathematics class. Teachers were expected to provide additional help filling out answer sheets and marking answers when necessary, such as transcribing answers to answer sheets for students who marked their answers in test booklets, and students were to have as much time as they needed to thoughtfully complete each day's set of problems. As noted earlier, accommodations such as additional time, frequent breaks from testing, assistance with answer sheets, and even altered font size are also available to students who need them in the general classroom. However, greater use is made of these accommodations in special education assessment settings.

The 2000–2001 academic year was the first year of mandatory participation for all students with disabilities in statewide assessment under IDEA. While eligibility criteria for the portfolio/rating scale Alternate Assessment for students with significant disabilities were well defined, no clear guidelines were available to assist IEP teams in determining which students would be best served by the modified test. IEP teams made test selection decisions on a student-by-student basis as part of the IEP development process. The state department of education restricted participation to students with disabilities who had scored below the 2.5th percentile on an existing norm-referenced standardized test of mathematics. However, the modified test's characteristics were not yet known and it was unclear whether that guidance would match students to the best assessment option. As noted earlier, participation in the Alternate Assessment amounted to less than 1% of the assessed population during that year. Participation in the modified assessment was approximately 1.5%.

A total of 570 fourth-grade students with disabilities from across the state were deemed eligible for the modified assessment by their IEP teams. In addition to requiring students to have either an IEP or a Section 504 plan, the eligibility criteria for the modified assessments stated that "The … team determines that the student is unable to take the general … assessment being considered" and "A preponder-

ance of data indicates that the student performs at or below the 2.5 percentile rank as measured by nationally and locally normed grade level measures of achievement in the academic area under consideration" (Kansas State Department of Education, 2000). This population consisted of students from every disability category except deaf–blindness, with the largest representation of students from specific learning disability, mental retardation, and noncategorical placement (Table 2).

These 570 students with disabilities took the 35-item modified assessment in four sessions as their only math achievement test. In addition to these students, 1,944 fourth graders taking the general assessment, including 182 students with disabilities, were randomly sampled by whole classes across the state for this research. These 182 students with disabilities represented nine categories of disability with the greatest number from speech/language impairment, specific learning disability, and noncategorical placement (Table 3).

Fifteen items from the modified assessment were assembled into a test booklet and administered to these students in their general education classrooms as an additional test session. These 15 items comprised the three new items written expressly for the modified test, the two previously unused items, and 10 items selected from among the four content standards areas, without duplicating an item on that year's general test form. Those 10 items included 1 unchanged item, 2 items with math simplification, and 7 items with language simplification. Except for the new items, the relative proportions of the different item types were quite similar to those of the modified test as a whole. The students in general education classes completed the extra test booklet before or after the four regular test sessions of the general assessment. In this way, the performance of special education and general

TABLE 2
Categories of Special Education Students Taking the Modified Assessment

Disability Category	*Frequency*	*% of Sample*
Hearing impairment	6	1.1
Visual impairment	1	.2
Speech/language impairment	9	1.6
Physical impairment	8	1.4
Specific learning disability	148	26.0
Emotional disorder	24	4.2
Mental retardation	139	24.4
Severe multiple disabilities	4	.7
Autism	11	1.9
Traumatic brain injury	1	.2
Noncategorical	140	24.6
Other health impairment	45	7.9
Incorrectly coded or unmarked	34	5.8
Total	570	100.0

TABLE 3
Categories of Special Education Students Taking 15 Modified Test Items in Addition to the General Assessment

Disability Category	*Frequency*	*% of Sample*
Hearing impairment	2	.1
Visual impairment	1	.1
Speech/language impairment	35	1.8
Physical impairment	3	.2
Specific learning disability	81	4.2
Emotional disorder	6	.3
Mental retardation	4	.2
Noncategorical	40	2.1
Other health impairment	8	.4
Incorrectly coded	2	.1
Total	182	9.5

education students on the modified test items was available for analysis and comparison.

Thus 2,514 students responded to the 15 modified items: 570 students who took only the 35-item modified assessment and 1,944 general and special education students who completed 15 extra items in addition to the 52-item general assessment form. The general assessment sample consisted of 963 females and 979 males (and two unmarked forms); 232 females and 338 males comprised the modified test group. Twenty-one percent of the students taking the modified test were African American compared to 8.5% of students in the general assessment, 7.7% were Hispanic compared to 5.3% on the general test, and 56% of the modified test group was White compared to 70% of the general assessment sample.

RESULTS

Initial assessment of test reliability showed good and comparable internal consistency for both tests (general test coefficient $\alpha = .87$, modified test coefficient $\alpha = .85$). Preliminary analyses were then conducted to evaluate whether the two tests measured the same construct for the two groups, a key step in evaluating validity. To assess unidimensionality and construct invariance for the two tests, each set of test items (the general test with 52 items plus the 15 extra items from the modified test; the modified test with 35 items) was subjected to confirmatory factor analysis (CFA) with a single-factor model using AMOS (Arbuckle, 1999). The 15 common items included two questions using the same item stem, so errors for these two questions were correlated in both models. Fit indexes for the CFAs are shown in Table 4. Both models found good support for unidimensionality with fit indexes of

TABLE 4
Fit Indexes for Confirmatory Factor Analysis of Single-Factor Model for Two Tests

Model	*GFI*	*AGFI*	*CFI*	*RMR*	*RMSEA*	χ^2	*df*	χ^2/df
General test: 67 items (52 items + 15 common items)	.94	.94	.90	.005	.019	3698.6	2143	1.726
Modified test: 35 items (including 15 common items)	.93	.92	.90	.009	.027	785.3	559	1.405

Note. GFI = Goodness-of-Fit Index; AGFI = Adjusted Goodness-of-Fit Index; CFI = Comparative Fit Index; RMR = Root Mean Residual; RMSEA = Root Mean Square Error of Approximation; *df* = degrees of freedom.

.90 or above and minimal residuals. Discussion and description of CFA fit indexes is available from several sources (see Byrne, 2001). Unidimensionality is a prerequisite assumption for the item response theory (IRT) analyses carried out in this study; thus this was a critical first step in demonstrating that the two tests measured the same underlying construct.

The content of the tests had previously been deemed comparable by the teacher teams who prepared the indicators to be assessed and the test items to measure those indicators but this assumption had not been empirically verified. CFA uses only statistical information about the covariance of measured variables (test items) and cannot address the content of the variables. However, because the common set of 15 items was included in both item sets, and both item sets showed good support for a single construct, this analysis substantiates the hypothesis that both sets of items measured the same construct.

To compare the functioning of the 15 common items, which were the only items that all students in both groups were exposed to, Mantel-Haenszel differential item functioning (DIF) analysis was conducted (Dorans & Holland, 1993). Evaluating the functioning of the common items was crucial in determining whether members of the two nonequivalent groups responded in similar fashion to the mathematics test items. DIF analysis showed no significant differences for 12 items; however, 3 remaining items had marginally significant DIF. In each case, the modified test group performed more poorly than expected. The presence of DIF does not necessarily imply bias but may reveal real differences in knowledge and performance between the two groups. Therefore, these three items were inspected to see whether the reason for the differential functioning could be identified. Two items dealing with geometry were on the cusp of significance. Both items used specific geometric vocabulary (shape names and geometric transformations). Although the modified test group performed slightly more poorly than expected relative to the

general test group, they still performed adequately on these two items, with *p* values of .66 and .45. The third item involved probability and was very difficult for the modified test group, with a *p* value of .25. In comparison, the next lowest *p* value on the modified test was .34.

One hypothesis for the significant DIF is different instruction for the two groups of students. The modified test group comprised the lowest performing students with disabilities, other than those eligible for the Alternate Assessment, and, according to the eligibility criteria for the modified test, all had prior individual math achievement test data showing them to be performing in the lowest 2% to 3% of their age groups. Due to this low math achievement, and signified by the fact that their IEP teams determined that they could not take the general assessment, they had most likely received their academic instruction in special education or resource room settings rather than in general math classes. It is likely that their instruction had focused more heavily on arithmetic calculation and functional math skills than on geometry or probability. Another possibility for different group response to these items is statistical. An examination of the two groups aggregated and distributed by total score on the 15 items reveals that there are fewer students in the general test group at total score levels where there are more students in the modified test, and vice versa. Because the mean scores on these items are so different for the two test groups and the group sizes are always smaller for one group than the other, sometimes one group is represented by an insufficient number of students at that total score. Estimates are less stable when sample sizes are low, so some of the apparent differences in item functioning could be due simply to the mismatch of the two total score distributions. Because the groups were known to be nonequivalent with respect to ability, as confirmed by their special education status and eligibility for either the modified or general test, and probably quite different in classroom instruction as a result of their special education placements in mathematics, the fact that three items showed marginally different functioning is not surprising. The other 12 items clearly measured the same thing for students in both groups. The DIF analysis provided additional, albeit qualified, support for the hypothesis that the 15 common test items measured the same constructs for both groups.

The main analysis used a one-parameter IRT model to place all 87 items from the general (52 items) and modified (35 items) assessments onto the same scale of item difficulty. Using the 15 common items as linking items, item difficulties were estimated for all 87 items. Then mathematics ability levels for all 2,514 students were estimated with the general assessment group as the reference group. In other words, the scale was based on the distribution of mathematical ability for members of the general assessment group and then ability for those of the modified assessment group was rescaled accordingly. According to convention, the student ability for the general assessment group was set to have a mean of 0 and a standard deviation of 1. In IRT, item difficulty and student ability are placed on the same scale so

that they can be compared directly. A student with math ability 1.0, for example, has a 50% probability of passing items with difficulty 1.0. Higher numbers represent more difficult items and higher levels of student math ability.

Figure 1 shows the result of IRT estimation of item difficulties for the two tests, with modified items shaded. General test items had a range of difficulty from –3.4 to 2.1 and a mean of –0.51. Modified test items ranged from –3.6 to –0.5 with a mean of –1.73, well below the overall item difficulty mean. As can be seen from the graph, modified items extend from about average difficulty to slightly below the easiest items on the general test, confirming that the modified test was of lower difficulty overall. Figure 1 also verifies that the modified test provided a larger number of lower difficulty items to which students with disabilities could respond. The general test provided 16 items of difficulty –1.0 or below while the modified test provided 28 such items.

The distribution of student mathematics ability is shown in Figure 2. Ability scores for students with and without disabilities who took the general assessment ranged from –2.86 to 2.58. Because they constituted the scaling reference group, their mean score was preset at 0 with a standard deviation of 1. Scores for students who took the modified test, all of whom had disabilities, ranged from –3.60 to 2.88 with a mean of –1.81, again much lower than the general assessment. Six students with disabilities were assigned spuriously high ability scores on the basis of having

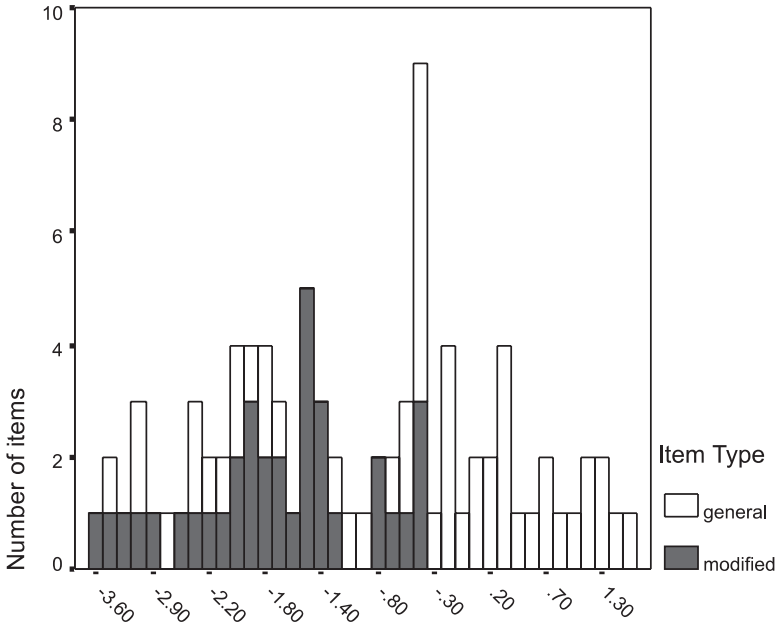

FIGURE 1 Numbers of items at various difficulty levels for both tests.

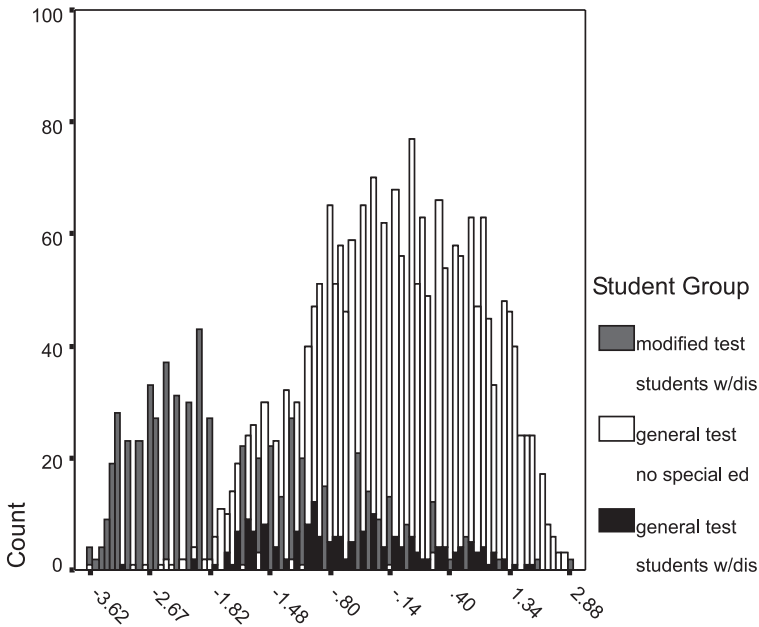

FIGURE 2 Distributions of mathematics ability estimates for students in three groups.

answered 33 or more of the 35 items on the modified test correctly. Their ability would most likely have been more accurately assessed with the general test, illustrating the importance of having an assessment with sufficient items at the appropriate level of difficulty. Alternatively, this problem might have been avoided if the modified test had included a few items of somewhat greater difficulty.

Because the two distributions of item difficulty and student ability use the same metric in IRT methodology, student ability can be directly compared with item difficulty. The modified test items, with difficulties ranging from –3.6 to –0.5 and a mean of –1.73, corresponded quite closely to the range of mathematics ability evidenced by most of the students who were deemed eligible for that test by their IEP teams (omitting the six incorrectly assigned students), which ranged from –3.6 to 0.69 with a mean of –1.81.

Next, test information was computed for each of the two tests to determine which test would provide more information about the achievement of students of different abilities. Test information at given ability levels was computed by summing the amount of information that each item in the test provides at that level and then plotting across the range of abilities. Because the one-parameter logistic model was used, for each item the amount of information equals pq, where p is the probability of passing the item at that ability level and $q = 1 - p$. The amount of in-

formation that an item can provide varies across different ability levels with maximum information obtained at the point where the item difficulty and the ability are the same. At this point an examinee will have a 50/50 chance of passing the item. Item information decreases as the examinee's probability of passing the item drops toward 0 or increases toward 1. Item and test information can be viewed as an index of how accurately a particular examinee at a given ability level can be measured. As such, test information can be used to decide which test among several is the appropriate one for a given sample of examinees.

In this case, the question of which test is more appropriate for which students can be addressed by comparing the test information curves for the two tests, shown in Figure 3. The range of appropriate ability levels assessed by each test is apparent, confirming the visual evidence that the modified test provides more appropriate items for students of lower math ability. Student math ability ranges can be compared with the test information curves to show that the modified test provides more information, and thus more appropriate math ability measurement, for lower performing students.

To facilitate the decision of the appropriateness of the two assessments, the relative efficiency of the modified and general tests was computed at each ability level and plotted over the ability range. The relative efficiency is the ratio of modified

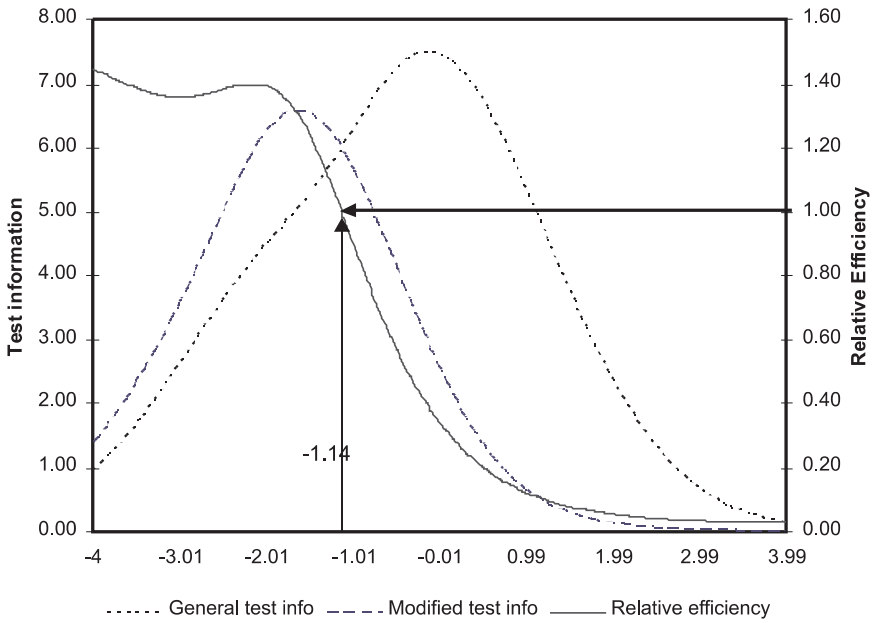

FIGURE 3 Test information curves for both tests and relative efficiency of the modified test compared to the general test showing the crossover point for ability.

test information to general test information at each ability level. The bigger the relative efficiency value, the more information the modified test can provide than the general test for that ability level. For the ability level where the value of the relative efficiency is 1.0, the two tests measure mathematics ability with equal precision. For lower ability levels the modified test is more accurate, whereas for higher abilities the general test is more accurate. The relative efficiency function, also shown in Figure 3, reveals that students with mathematics ability of less than –1.14 are more reliably assessed with the modified test. The expected score of an individual at math ability level –1.14 can be computed by summing the probabilities of passing each item on each test at that ability level. For comparison, on the general test this expected score is about 20 items out of 52. A student at the same math ability would be expected to earn a score of 22 out of the 35 items on the modified test.

Of course, a student's math ability is not known beforehand. Two approaches can be taken to resolve this problem. One is to estimate the student's ability using performance on a test taken in a previous year using IRT ability estimation procedures from the previous test group. The second is to estimate an approximate percentage of students who should take the modified test from a representative sample of the student population of the state. For the current general assessment sample, this percentage would include approximately 11.4% of students using the cutoff point of mathematics ability level –1.14. It is important to note that for this second approach these percentages apply only to this particular pair of tests and cannot generalize to any other assessments.

DISCUSSION

The charge to develop and administer a specialized mathematics assessment with built-in modifications for a particular population of students with disabilities afforded the opportunity to use IRT methodology to compare the new test to the general mathematics assessment. Both tests conform to a common unidimensional model of mathematics as an overall construct. Each test assesses a single major factor, suggesting that the two tests assess the same construct domain, namely mathematics achievement, for the two populations of students. Comparison of student performance on the 15 common items using DIF methodology provides additional, qualified support for this conclusion with 3 items showing marginal differential functioning. Both tests had adequate reliabilities for their intended populations. All eligible students with disabilities were administered the modified test, so the entire population of responses to these items was available. A sufficiently large sample of students in general education, with special needs and without, was administered the anchor block of items, enabling all items on both tests to be compared on a single scale of difficulty.

The results demonstrate that an assessment measuring the general curriculum with reduced difficulty could be developed for students with disabilities for whom the general assessment would not be a valid test. The item difficulty figures confirmed that a major objective of the modified test had been accomplished, which was to provide more lower level items and omit too-difficult items while maintaining the overall curricular coverage of the general mathematics assessment. More accurate measurement was achieved even though the modified test was constrained to fewer items.

Alignment with the high quality of the general curricular standards (Finn & Petrilli, 2000) was maintained through a process including review and revision of assessed instructional indicators from the general curricular standards to prepare them for use by students with disabilities; analysis and alteration of existing field-tested items, which had been written for the general standards, for concordance with the modified indicators; and the preparation of only three new items where existing items were not adaptable to the modified test. At each step, statewide teacher teams consisting of grade-level mathematics and special education teachers performed the actual work with state department of education and test contractor personnel supervising. Items prepared to assess the modified indicators were field tested with the target population before final selection for each of the grade-level modified test forms. The same table of specifications was used for each modified test as for its corresponding grade-level general test. The test development procedure was identical to that followed for the general state assessments.

A second major objective was to maintain standardized administration across the state while providing appropriate modifications for students with disabilities. Defining and building in predetermined modifications meant that students with disabilities who were eligible for the modified test were assessed under common conditions and their performance could be compared within the same year or across years. Individualized and poorly controlled modifications to testing were not permitted and all other changes in the test environment were considered accommodations that did not affect the measured construct.

The results also provide an estimate of the proportion of lower achieving students for whom our state's modified test might provide more appropriate measurement than the general assessment. In this study, the lowest 11.4% of the general assessment sample distribution of math ability could be measured more precisely by the modified test. That such a large proportion of students fall in this range is directly due to the rigor of the general assessment in our state, which contains items of considerable difficulty even for nondisabled students. This large a proportion probably does not represent the number of students who *should* take a modified test instead of the general assessment. It does demonstrate that, for these two tests and this sample of the general population of students, the simpler modified test provides more information about the performance of students below this level of mathematics achievement in terms of providing sufficient items of the appropriate difficulty. These estimates are only meant to compare these two assessment instru-

ments; they do not address the relationship of the student's instructional curriculum to the items on either assessment. That is the charge of the IEP team when making individualized instructional and assessment decisions.

The most significant limitation of this study is apparent from the previous discussion, which is that these results are applicable only to the two specific mathematics assessments devised for fourth-grade students in our state; no generalization to other instruments, populations, or content areas can be made. A similar set of analyses would have to be performed for each pair of tests, at each grade level and for each content area, for which a grade-level specific modified assessment was planned. In addition, the modified test did not have as many items as the general test, which is a beneficial feature for assessing students who need more time per item but a potential problem with content coverage, even though the items were selected to measure the same instructional indicators according to the same table of specifications. Furthermore, three items did show some differential functioning for the two groups, although this is not a surprising outcome given the considerable ability differences between the groups and the probable variation in their math instruction.

The development of a new test comes at considerable cost, as test developers are well aware. The finding that more than 10% of the students in this study would have been more accurately measured on the modified test raises the question of how these lower performing students could validly be assessed other than with a separate but linked standardized instrument. The mathematics ability of a greater number of lower achieving students would be more accurately measured if the general assessment included more items of lower difficulty. A selection from the 87 items making up this pair of assessments could probably be made to form a test that would measure almost all of these students equally well, although the test would be longer than the modified test and there would likely be unnecessary items, either too easy or too hard, for each student. A more appropriate single assessment using universal test design concepts (Thompson, Johnstone, & Thurlow, 2002) might be developed using techniques like those used here, with teacher teams including special educators and teachers of English as a Second Language to address the needs of special populations who will take the test. Another alternative is computer adaptive testing, which is designed to provide sufficient items at each level of difficulty so that each student is exposed to the minimum number of items needed for an accurate estimate of ability. However, computer adaptive testing requires a level of expertise and availability of technology that may put it outside the reach of many state testing programs.

Beyond the statistical properties of tests, there are pressing policy issues that must also be used to guide assessment development and decision making. The alignment of instructional curriculum with test content is a critical validity issue with respect to our current demand for accountability assessment. The IDEA 1997 amendments, in mandating that students with special needs be included in district- and statewide testing, were driving toward just this type of alignment with the goal that special needs students be exposed to the general curriculum to compete on

large-scale assessments. NCLB continues this accountability thrust with its increased demands for large-scale assessment and the requirement that all students be assessed at more grade levels.

This study demonstrated that grade-level curricular standards can be assessed in a way that allows students with disabilities to more accurately demonstrate their knowledge. This is a different situation than off-grade testing using tests that have been vertically equated across grade levels, which may actually measure different content from grade to grade. Off-grade testing does not respond to NCLB's call for assessment in the general curriculum, and may in fact disadvantage students with disabilities who would be tested on content that should not be part of their grade-level curricula. Students with disabilities should not automatically be instructed on lower grade level material, although they may well need simplification and support within the general curriculum.

The tension between accessibility and lofty standards is high when assessments are deliberately *not* "dumbed down" to allow most students to achieve passing scores but are intentionally rigorous and broad in scope. High standards often result in tests on which few students obtain outstanding scores and many achieve only minimal or basic proficiency. This investigation shows that students with disabilities who perform within the lowest few percentile ranks of the achievement distribution are truly not being equitably assessed by the general assessments. To be reasonably assessed in a manner that dignifies their individual achievement goals and provides real information about their progress and the efforts of their schools and districts, assessments such as the modified test studied in this project must be developed, used, and evaluated. While limiting these tests to a small percentage of students is appropriate, this research has explored one potential method of meeting this need.

This project was intended to evaluate whether a modified instrument could be devised that corresponded to the content of the general test in a manner that would allow lower achieving students to demonstrate their knowledge of mathematics content, as well as to provide information about which students could be assigned to that test instead of the general assessment. Further study is needed with other populations, other grade levels, and other content areas to determine whether modified assessment options meeting the IDEA 1997 guidelines and NCLB can be developed on a broader scale. Such assessments could fulfill the letter and the spirit of federal law requiring inclusion of all students in grade-level assessments while providing a more valid measurement of student and school progress within the general curriculum.

ACKNOWLEDGMENTS

This study was supported by funding provided by the Kansas State Department of Education. Opinions expressed herein are those of the authors and do not necessarily reflect those of the sponsoring agency.

REFERENCES

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report on language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Arbuckle, J. L. (1999). AMOS [computer program]. Chicago: SmallWaters.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Destefano, L., Shriner, J. G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessment. *Exceptional Children, 68,* 7–22.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Finn, C. E., & Petrilli, M. J. (Eds.). (2000). *The state of state standards 2000.* Retrieved January 12, 2005, from the Thomas B. Fordham Foundation Web site http://www.edexcellence.net/doc/Standards2000.pdf

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teacher's knowledge of accommodations as a validity issue in high-stakes testing. *Journal of Special Education, 32,* 175–183.

Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 105–17, 37 Stat. 111 (1997).

Johnson, E., Kimball, K., Brown, S. O., & Anderson, D. (2001). A statewide review of the use of accommodations in large-scale, high-stakes assessments. *Exceptional Children, 67,* 251–264.

Johnson, E., & Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assessment for Effective Intervention, 29*(3), 35–45.

Kansas State Department of Education. (2000). *Kansas assessments with modifications: Eligibility criteria and overview of the Kansas assessments with modifications for 2000–2001 academic year.* Topeka, KS: Author.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Schulte, A. A. G., Elliott, S. N., & Kratochwill, T. R. (2001). Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performances of students with and without disabilities. *School Psychology Review, 30,* 527–547.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. R., & Poggio, J. P. (2002). *The differential impact of accommodations in statewide assessment.* Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. R., & Poggio, J. P. (2003). The differential impact of accommodations in statewide assessment: Research summary. Retrieved January 12, 2005, from http://education.umn.edu/NCEO/TopicAreas/Accommodations/Kansas.htm

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (NCEO Synthesis Rep. 44). Retrieved January 12, 2005, from http://education.umn.edu/nceo/OnlinePubs/Synthesis44.html

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (1998). *Testing students with disabilities: Practical strategies for complying with district and state requirements*. Thousand Oaks, CA: Corwin Press.

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations.* Retrieved January 12, 2005, from http://www.ihdi.uky.edu/msrrc/PDF/Tindal&Fuchs.PDF