Differential Item Functioning Comparisons on a Performance-Based Alternate Assessment for Students with Severe Cognitive Impairments, Autism and Orthopedic Impairments

Cara Cahalan Laitusis
Behroz Maneckshana
Lora Monfils

Educational Testing Service

Lynn Ahlgrim-Delzell

University of North Carolina at Charlotte

Abstract

The purpose of this study was to examine Differential Item Functioning (DIF) by disability groups on an on-demand performance assessment for students with severe cognitive impairments. Researchers examined the presence of DIF for two comparisons. One comparison involved students with severe cognitive impairments who served as the reference group and students with autism and severe cognitive impairments who served as the focal group. The other comparison compared students with severe cognitive impairments (reference group) and students with severe cognitive impairments and orthopedic impairments (focal group). Results indicated a moderate amount of DIF for the autism comparison and a negligible amount of DIF for the orthopedic impairment comparison. In addition researchers coded all test items based on characteristics likely to favor one of the three groups. Although several of the hypothesized coding categories resulted in accurate prediction of DIF, the study was limited to items from one testing program for students in one state. More research is needed to see if these hypotheses can be replicated across testing programs and populations.

Introduction

Background

Recent federal legislation, such as the No Child Left Behind Act of 2001 (2002) and the reauthorization of the Individuals with Disabilities Education Act (1997), have redefined the role of the federal government in K-12 education. According to NCLB, each state must define adequate yearly progress (AYP) and provide measurable assessment of student performance. Along with mandating annual student testing in grades 3-8, NCLB stipulates that assessments provide accommodations for students with disabilities as defined in the Individuals with Disabilities Education Act (IDEA, 1997 and 2004). In addition, NCLB mandates the reporting of assessment results and AYP by student groups based on poverty, race and ethnicity, disability, and limited English proficiency. All of these mandates have changed the landscape of state assessments, particularly for students with severe cognitive disabilities who were often exempt from taking state assessments prior to NCLB.

Along with the increase in the level of participation in assessments of students with severe cognitive disabilities, the format and content of alternate assessments have changed in recent years. The majority of states now rely on a portfolio or body of evidence analysis to assess the adequate yearly progress (AYP) of students with severe cognitive disabilities. Other states have employed rating scales or checklists, analysis of the student's Individualized Education Program (IEP), or other assessment formats such as the use of performance assessments to assess AYP (Thompson, Johnstone, Thurlow, & Altman, 2005; Roeber, 2002). The content of these alternate assessments is based on modified state academic achievement standards in reading, mathematics, and in some cases, science.

Research on the psychometric quality of state assessments for students with disabilities is quite limited. Nearly all of the existing psychometric research has focused on the impact of test accommodations for students with mild to moderate disabilities or on the use of portfolio assessments for students with severe cognitive disabilities (see Sireci, Scarpati, & Li; 2003; Pitoniak & Royer, 2001). Some research has used operational test data to examine Differential Item Functioning (DIF) between test takers with and without disabilities or test takers that receive testing accommodations and those that take a standard administration (see Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Bolt & Bielisnki, 2002; Bolt & Yesseldyke, 2006; Cahalan-Laitusis, Cook, & Aicher, 2004; and Koretz & Hamilton, 2000; Pitoniak, Cook, & Laitusis, 2006). To date, none of the published DIF research has focused on alternate assessments for students with severe cognitive impairments. The purpose of this study is to examine Differential Item Functioning by disability groups on an on-demand performance assessment for students with severe cognitive impairments.

Differential Item Functioning

An item is said to function differently for two or more groups if the probability of answering an item correctly is a function of group membership for examinees of the same ability (Camilli, 1993); that is, the item is harder (or easier) for equally able examinees belonging to one group versus another comparison group. In order to distinguish between DIF and item impact, DIF detection methods require that examinees from comparison groups be matched on the primary or essential underlying ability being measured (Shealy & Stout, 1993a, 1993b). Statistics of item impact do not require that this matching be done. When statistical methods flag an item as showing DIF, the item is commonly subjected to judgmental review to determine whether causes of differential difficulty are relevant to the construct being measured. Sources of differential difficulty may involve the presence of abilities secondary to the construct the item is designed to measure, such as when test items tap multiple proficiencies and these proficiencies systematically vary across groups of examinees (Camilli, 1992), differences in the instructional backgrounds of students (Muthen et al., 1995), or simply how some features of a test item interact with examinee characteristics.

Method

Differential Item Functioning is not commonly included in the analysis of results from state alternate assessments due to small sample sizes, prevalent use of portfolio assessments (where DIF is difficult to calculate), and the level of heterogeneity of the relevant test- taker population. However, in this study, data from a typical state alternate assessment program which has a large and diverse population with significant cognitive disabilities were used across two content areas; English-language arts (ELA) and mathematics (Math).

The primary purpose of this study was to ascertain if certain item characteristics result in DIF for a specific disability category. To determine if specific item characteristics impact the performance of students with specific disabilities (mental retardation, autism, and orthopedic impairments), each item was coded based on a set of characteristics that are hypothesized to more likely advantage individuals from one subgroup over another. For example, it is likely that items that require memorization of facts will be relatively easier for students with autism than for students with mental retardation who have no other disability classification. The diagnostic manual used by most clinicians (American Psychiatric Association, 2000) specifies that for individuals with autistic disorder "tasks involving long-term memory (e.g., train timetables, historical dates, chemical formulas, or recall of the exact words of songs heard year before) may be excellent" (p. 68) while long term memory of facts is not included in the diagnostic criteria for individuals with mental retardation.

Using DSM-IV diagnostic criteria, as well as prior knowledge of research coupled with experience teaching individuals with disabilities, a set of item characteristics codes was developed and reviewed by three researchers familiar with each of three disabilities (i.e., mental retardation, autism, and orthopedic impairments). Then two experts on students with disabilities coded all items based on these characteristics. Finally test items were examined to determine if those that were predicted to have DIF (based on the coded characteristics) were in fact more likely to actually exhibit DIF. The next section provides detailed information on the measures, samples of test takers, DIF procedures, and item classification codes.

Measures

The test in the study is an on-demand performance assessment that serves as an alternate assessment for the standardized statewide assessment test. Eligibility for the alternate assessment is based on the student's Individualized Education Program (IEP), which typically reflects an emphasis on functional life skills. The decision to participate in the alternate assessment is made only if the student cannot take the standardized statewide assessment even with accommodations. The alternate assessment can only be administered by certificated or licensed school staff members who have completed training in administering the exam. Preferably, the examiner should be the special education teacher or case carrier who regularly works with the students. Two content areas: English language arts and Mathematics were used for these analyses. The test content was constructed in degrees of difficulty according to specified levels.

Level I is for those students in grades 2 – 11 with the most significant cognitive disabilities who are functioning at or below the developmental level of 24 months. Levels II-V are for specific grade level bands. Level II is for grades 2 and 3, Level III is for grades 4 and 5, Level IV is for grades 6, 7, and 8, and Level V is for grades 9, 10, and 11. For students who are in ungraded classes, but are developmentally functioning above 24 months, the following formula is used to determine their appropriate grade: Grade= Age – 5. Table 1 below describes the levels, grades, and content areas for the alternate assessment.

At each level, 10 items are administered (8 operational test items and 2 field-test items) per content area (ELA and Math). Although more items would be desirable to hopefully increase the reliability of scores; the existing reliability was sufficient to use the set of items as a matching criterion (see results section for details on reliability of total test scores for ELA and Math). For Level I, each item is graded on a 0-5 point range, whereas for Levels II to V, each item is graded on a 0-4 point range. Total raw scores were converted to scale scores which are reported on a 15 to 60 point scale.

Table 1

Test Levels by Grade and Content Areas

Test Levels	Grades	Content Areas	
I	2-11*	ELA	Math
II	2-3	ELA	Math
III	4-5	ELA	Math
IV	6-8	ELA	Math
V	9-11	ELA	Math

Note. *The Level I assessment is for students who are functioning at or below the developmental level of 24 months, regardless of grade level or chronological age. ELA=English language arts, Math=Mathematics

Sample

Thirteen different disability subgroups were identified in the testing population, but sample sizes limited this study to only three disability subgroups[1]. These subgroups included students with: 1) Mental Retardation; 2) Orthopedic Impairment with Severe Cognitive Impairment; and, 3) Autism with Severe Cognitive Impairment. To reduce the impact of additional factors, we excluded Out-of-level test takers and English-language learners from the sample.

Students were matched in terms of their abilities (as measured by their total scale score). In most prior research studies of DIF, the reference group has been defined as students without disabilities. Since alternate assessments are not taken by students without disabilities, we lacked a clear reference group for this study. A decision was made to consider students with Mental

[1] The minimum sample size for the disability groups was 200 in the focal group and 600 across both the focal and reference groups.

Retardation and no other disability (MR) as the reference group because it comprised the largest group of test takers and it was assumed that all students in the sample had some form of mental retardation (or severe cognitive disability). The focal groups were defined by their unique disability classification (in addition to their severe cognitive impairment). These two focal groups were: 1) Autism with Severe Cognitive Impairment (AU); and, 2) Orthopedic Impairment with Severe Cognitive Impairment (OI). Table 2 below shows the number of students in the focal and reference groups that are the focus of this study.

Analysis Procedures

Prior to conducting DIF analyses, exploratory factor analyses were conducted to ensure that the total test score used as a matching criterion was measuring a similar construct or constructs for each group. A comparison of the size of the eigenvalue for the first factor with the sizes of the eigenvalues for the subsequent factors supported unidimensionality for all groups. These analyses indicated that the tests are essentially unidimensional and provide evidence that the ELA and Math tests are each measuring a single construct (Maneckshana, Laitusis, & Monfils, 2006). The next step was to examine the items for differential item functioning (DIF). An item is said to display DIF if examinees from different groups have differing likelihoods of success on the item after matching on the ability the test is intended to measure. Although there are several methods of evaluating DIF, the method used for this paper involved the Standardized Mean Index, (Dorans &, Schmitt, 1991, Zwick, Donoghue, & Grima, 1993) along with the Mantel chi-square statistic (Mantel, 1963).

Table 2

Sample Size by Reference/Focal Comparison, Level, and Content Area

Level	Group	ELA		Math	
		MR/OI Comparison	MR/AU Comparison	MR/OI Comparison	MR/AU Comparison
I	Reference (MR)	1848	1848	1845	1845
	Focal	1375	1048	1375	1047
	Total	3223	2896	3220	2892
II	Reference (MR)	1302	1302	1304	1304
	Focal	204	927	205	926
	Total	1506	2229	1509	2230
III	Reference (MR)	1592	1592	1592	1592
	Focal	230	915	231	916
	Total	1822	2507	1823	2508
IV	Reference (MR)	2951	2951	2958	2958
	Focal	360	1198	360	1200
	Total	3311	4149	3318	4158
V	Reference (MR)	3338	3338	3336	3336
	Focal	332	762	331	760
	Total	3670	4100	3667	4096

For polytomous item types such as those used in this alternate assessment, the Mantel chi-square statistic with one degree of freedom is typically used in conjunction with the standardized mean difference (SMD) method to assess DIF. The Mantel chi-square test is a DIF procedure appropriate for use with items with ordered response categories. This procedure compares the item means of the two groups matched on the test's total score. Since all the items in this study were polytomous in nature, the standardization procedure in conjunction with the Mantel chi-square statistic was used to classifying items into three categories: Category A (negligible DIF), Category B (slight to moderate DIF), and Category C (moderate to large DIF)[2].

Item Coding

To determine if specific item characteristics impact the performance of students with mental retardation (MR), severe cognitive impairment with autism (AU), and severe cognitive impairment with orthopedic impairments (OI), each item was coded prior to examining DIF. Items were coded based on the state standard(s) the item was measuring as well as other characteristics that were related to the students' disability classification. The trait-related classifications were developed by three special education researchers and were based on prior research (see appendix for complete list of literature) and the classification criteria for the Mental Retardation and Autistic Disorder (American Psychiatric Association, 2000). Each category was

[2] The flagging criteria for polytomous items are as follows: C items are those items that have a Mantel Chi-square p-value < 0.05 and a |SMD/SD| > 0.25; B items are items that have a Mantel Chi-square p-value < 0.05 and a |SMD/SD| > 0.17, and the remaining items are classified as A. SMD is the Standardized Mean Difference index, and SD is the total group standard deviation of the item score. A negative SMD value shows that the question is more difficult for the focal group whereas a positive value indicates that it is more difficult for the reference group.

developed using prior research and hypothesized to result in DIF favoring (1) MR compared to AU, (2) AU compared to MR, (3) MR compared to OI, and (4) OI compared to MR.

Two special education researchers coded the 100 items (50 Math and 50 ELA) used in this study based on 15 overarching item categories. In both Math and ELA, there were 10 items per level and 5 levels, resulting in a total of 50 items for the study. Two of these coding categories (short term memory and attention deficits) were hypothesized to result in DIF favoring both the OI and AU focal groups when compared to the MR reference group). In addition, several categories had subcategories. For example the social exchange (SE) category had three more specific classifications involving social play (SE1), role play (SE2), and social events (SE3). Multiple coding categories could be applied to the same item and in many cases this is what occurred. Of the 15 overarching categories, 10 were hypothesized to impact the MR-AU comparison; these included the following:

- Attention deficits
- Imitation
- Joint Focus of Attention
- Perseveration
- Rote Learning
- Social Exchange
- Short Term Memory
- Symbolic Language/ Play
- Theory of Mind
- Visual Spatial Orientation

Seven of the 15 categories were hypothesized to impact the MR-OI comparison; these included the following:

- Attention deficits
- Fine Motor
- Generalization
- Proprioception
- Receptive/ expressive language
- Short Term Memory
- Visual processing

Inter-rater agreement. Overall initial inter-rater agreement across all coding categories and all items was 92% (91% for ELA and 92% for Mathematics). Agreement was calculated for all coding categories separately and initial inter-rater agreement by coding category ranged from 74% to 100%. Disagreements were resolved through discussion between the two raters. Several of the coding categories with lower levels of agreement (ranging from 74% to 93% initial agreement) were revised to more accurately reflect the intentions of the researchers and to increase inter-rater agreement. These categories included visual-spatial orientation, social exchange, rote learning, and perseveration.

Results

Preliminary analyses were done to examine differences in means and standard deviations and in the reliability of the scores for each of the disability groups (i.e., the Autism, Mental Retardation, and Orthopedic Impairment groups). The means, standard deviations, reliabilities, and standard error of measurement (SEM) for each group (MR, AU, and OI) by level (I-V) are displayed in Tables 3 and 4, for ELA and Math, respectively.

Mean Scores and Variability of Scores

Level I showed the most variability and the largest differences in means across the three groups. These differences could be attributed to the wider age range of students included in Level I (all students included from grades 2 to 11 functioning below 24 months developmentally). When comparing group means within each level, for ELA and Math, the means of the Autism group are higher than the means of other groups, at the lower levels (Levels I and II), but lower at the upper levels, particularly at Levels III and IV. The Orthopedic Impairment group mean was lowest at Level I but improved when looking at upper levels in both ELA and Math. The mental retardation group had the highest mean at Level V.

Reliability

Score reliability was estimated using coefficient alpha. The higher the reliability coefficient for a set of scores, the more likely individuals would be to obtain similar scores if they were retested. The reliability estimates for ELA range from .86 to .94 for the three groups, showing high consistency. The reliability estimates for Math, ranging from .87 to .92, also show high consistency for all three groups.

Standard Error of Measurement

The squared standard error of measurement (SEM) is an estimate of error score variance. The spread of SEMs for the groups below are moderate. As seen in Tables 3 and Table 4, the largest standard error of measurement (SEMs) for ELA for all groups was at Level I. Level II had the smallest SEMs. For Math the largest SEMs for all groups were at Level V.

Table 3

Scale Score Summary Statistics: English-language Arts

Level	Subgroup	N	Mean	SD	Reliability	SEM
I	MR	1848	45.81	12.96	.92	3.68
	OI	1375	38.65	12.43	.90	3.96
	AU	1048	46.49	10.52	.86	3.91
II	MR	1302	36.35	7.90	.89	2.59
	OI	204	36.86	8.49	.90	2.72
	AU	927	36.75	9.39	.91	2.81
III	MR	1592	35.66	9.60	.90	3.08
	OI	230	37.48	10.52	.91	3.23
	AU	915	34.25	10.20	.92	2.97
IV	MR	2951	34.86	9.16	.89	2.99
	OI	360	35.86	9.82	.90	3.15
	AU	1198	33.49	10.69	.92	3.11
V	MR	3338	37.41	10.30	.91	3.03
	OI	332	36.36	11.27	.93	3.03
	AU	762	35.52	11.97	.94	3.01

Table 4

Scale Score Summary Statistics: Mathematics

Level	Subgroup	N	Mean	SD	Reliability	SEM
I	MR	1845	34.94	11.21	.92	3.18
	OI	1375	28.85	10.14	.92	2.84
	AU	1047	35.94	8.99	.84	3.55
II	MR	1304	38.24	7.98	.87	2.88
	OI	205	39.41	9.13	.89	3.09
	AU	926	38.26	8.91	.87	3.16
III	MR	1592	39.02	10.03	.89	3.30
	OI	231	39.58	11.08	.92	3.17
	AU	916	38.06	11.26	.91	3.40
IV	MR	2958	33.15	9.75	.88	2.94
	OI	360	33.39	10.79	.90	2.83
	AU	1200	32.88	10.56	.88	3.02
V	MR	3336	33.92	9.09	.90	3.37
	OI	331	32.28	9.46	.91	3.37
	AU	760	32.67	11.00	.92	3.59

DIF Results

Tables 5 (ELA) and 6 (Math) are summaries of the items by DIF category for all levels and comparisons. Both tables list the numbers of items in each of three DIF categories (A=negligible, B=slight to moderate, and C=moderate to large); as well as the direction of the DIF (a positives value indicates that the item is more difficult for the reference group and a negative value indicates that the item is more difficult for the focal group). Each level included at least 8 test items but some levels also included 2 field test items when sample size permitted. When field test sample size did not permit a DIF analysis to be done, the field test items are listed under the not applicable (N/A) category.

For the MR/Orthopedic Impairment (MR/OI) comparisons, most items (n=77) exhibited negligible DIF (category A); 7 items exhibited slight to moderate DIF (category B), and 2 items exhibited moderate to large DIF (category C). DIF analyses could not be calculated on 14 field test items due to low sample sizes in the OI group. The two items that exhibited category C DIF included a Level I ELA item which was flagged as C+ and a Level II Math item which was flagged as C-. Of the 7 items flagged at B level DIF, there was no clear patter in terms of the content area, direction of DIF, or level of the assessment.

For the MR/Autism (MR/AU) comparisons there was substantially more items flagged as exhibiting DIF than was found for the MR/OI comparisons. Analyses could not be completed for 9 of the field test items due to low sample sizes. Of the remaining 91 items included in the MR/AU comparison across levels and content areas, 58 items were classified as negligible DIF (category A). Of the remaining items, 23 exhibited C-DIF and 10 exhibited B-DIF. The majority of the items exhibiting B or C level DIF were found on the ELA assessment rather than the mathematics assessment. On both assessments the direction of the DIF (positive or negative) was fairly balanced, but the largest proportion of items flagged for B or C level DIF were found at Levels II, III and IV with fewer items B and C level DIF items found in Levels I and V.

Table 5

Operational and Field Test Items by DIF Category for English Language Arts

	DIF Categories					
Level	C+	B+	A	B-	C-	N/A
			MR/AU			
I	0	0	7 (2)	1	0	0
II	3 (1)	0	2	1	2	0 (1)
III	3	0	2	1	2	0 (2)
IV	1 (1)	0 (1)	5	0	2	0
V	0	1	4 (2)	1	2	0
			MR/OI			
I	1	0	7 (2)	0	0	0
II	0	1	7 (2)	0	0	0
III	0	1	6	1	0	0 (2)
IV	0	0	8	0	0	0 (2)
V	0	0	8	0	0	0 (2)

Note. Field Test Items are in parentheses

Table 6

Operational and Field Test Items by DIF Category for Mathematics

	DIF Categories					
Level	C+	B+	A	B-	C-	N/A
			MR/AU			
I	0	0	8 (2)	0	0	0
II	1	1	4	0	2	0 (2)
III	1	0	5	1	1	0 (2)
IV	0	1	6 (1)	1	0 (1)	0
V	0	0	8	0	0	0 (2)
			MR/OI			
I	0	0	8 (2)	0	0	0
II	0	1	5	1	1	0 (2)
III	0	1	7	0	0	0 (2)
IV	0	0	8	0	0	0 (2)
V	0	1	7	0	0	0 (2)

Note. Field Test Items are in parentheses

Observed DIF by Content Strands

In general, for ELA the Autism group showed positive DIF for items covering content areas such as sight word reading, writing strategies, and reading/word analyses. Items covering content areas such as reading/reading comprehension, listening and speaking favor the mental retardation group. For Math the Autism group showed positive DIF for content areas such as Measurement and Geometry. Positive DIF favoring the mental retardation group covered content areas such as Number Sense and Data Analysis and Statistics. This pattern, however, was not consistent across all Math levels.

Fewer items showed DIF for the MR/OI reference-focal group comparison than the MR/AU comparison. Among the few items that did show DIF, items covering content areas such as listening and speaking applications favored the OI group. In addition, 2 items at Level 3 covering the topics of writing/writing strategies favored the MR group.

Observed DIF by Hypothesized Coding Categories

Coding categories for the Orthopedic Impediment group were not analyzed because only 2 items showed C-level DIF for this comparison (1 item for ELA and 1 item for Math). Full descriptions of the MR/AU item characteristics coding categories are displayed in Table 7. It is important to note that Table 7 includes hypothesized causes of DIF and not the outcomes of the actual DIF analyses.

Table 7

Hypothesized Coding Categories

		Advantage Autism
Visual-Spatial Orientation	VS1	Items in which the answer can be found within the visual stimuli of the test materials.
	VS2	Items that involve sorting, matching, or recognizing patterns.
Rote Learning	RO1	Items that rely primarily on rote learning other than phonics (e.g., counting, sight words).
	RO2	Items that rely on rote learning of phonics to decode unfamiliar words (not functional sight word).
Attention Deficits	AT	Items that require longer duration of attention to task (includes multiple questions within an item).
Short Tem Memory	STM	Items that require remembering information before answering (e.g., multiple step directions).
Sequencing	SQ	Items that involve sequencing (e.g., which number comes before 2).
		Advantage MR
Visual-Spatial Orientation	VS3	Items without any visual stimuli (items that are given verbally only).
Social Exchange	SE1	Items which require a social exchange or social play (e.g., short conversation, taking turns).
	SE2	Items involving symbolic play or role play.
	SE3	Items that relate to social events (playing with friends, celebrating holidays).
Symbolic Language/Play	SU1	Items that use figurative vs. literal language.
	SU2	Items focusing on comprehension versus word recognition in reading.
Imitation	IM*	Items that require learning a new motor/ verbal imitation during the task.*
Theory of Mind	TM1	Items requiring the student to understand another's point of view
	TM2	Items that use first or second person pronouns
Perseveration	PS1	Items that involve reuse of the same materials in a different way (e.g., using the same manipulative to ask multiple questions within the same item)
Joint Focus of Attention	JA1*	Items that require drawing another's attention to the task or one's work.*
	JA2	Items that focus on listening comprehension

Note. An asterisk (*) indicates that none of the items were coded as having this characteristic.

Table 8 includes a summary of the number of items by code and content area that are hypothesized as advantaging the groups with autism and mental retardation.

Table 8

Number of items by Hypothesized Code and Content

		Number of Items	
Advantage	Characteristic	ELA	Math
Autism	VS1	14	3
	VS2	3	13
	RO1	15	10
	RO2	3	0
	AT	13	8
	STM2	3	10
	SQ	1	2

Table 8

Number of items by Hypothesized Code and Content (continued)

		Number of Items	
Advantage	Characteristic	ELA	Math
Mental Retardation	VS3	5	2
	SE1	12	4
	SE2	4	1
	SE3	0	1
	SU1	0	3
	SU2	0	3
	TM1	4	3
	TM2	39	19
	PS1	1	14
	JA2	2	6

Characteristics favoring AU on ELA

Only one coding characteristic, (rote learning of phonics to decode unfamiliar words, RO2), consistently identified items with DIF favoring students with autism. In all cases, items that were coded as RO2 had DIF. In addition the items which combined decoding with sight words exhibited less DIF than items that required only decoding without sight word recognition. Another coding category, rote learning (RO1), appeared to show some promise in identifying items with DIF. Slightly less than half (7 of 15) of the ELA items coded as primarily relying on rote learning (RO1) showed DIF favoring students with Autism. Two other coding categories (AT and VS2) were used frequently but did not appear to be related to hypothesized DIF. Several (5 of 12) ELA items that required longer duration of attention to tasks (AT) showed DIF favoring students with autism. Nearly all of the AT items that showed DIF, however, were coded as having a combination of AT with RO1. All four ELA items coded as AT and RO1 showed DIF favoring students with autism. Items that required finding the correct answer within visual stimuli (VS1) did not appear to result in an increase in DIF (2 of 12 items with this characteristic exhibited DIF).

Characteristics favoring AU on Mathematics

Only two mathematics items showed DIF favoring students with autism. None of the coding categories appeared to be related to these two items. Four coding categories, rote learning (RO1), attending to multiple tasks (AT), short term memory (STM2) and sorting, matching, and recognizing visual patterns (VS2) were used frequently enough (7 to 12 items) to show that these characteristics on their own do not appear to be responsible for DIF. For example the two items that did exhibit DIF favoring students with autism were both coded as RO1, however 5 other items also coded as RO1 did not exhibit any DIF. The other frequently coded characteristics (AT, STM2, and VS2 with 8, 9, and 12 items respectively) were not used for any items that showed DIF.

Characteristics favoring MR on ELA

All five items that had a combination of three coding characteristics; verbal only (VS3), required a social exchange (SE1), and used first or second person pronouns (TM2) showed DIF favoring students with mental retardation. The use of first or second person pronouns was used in many other ELA items (36 items) and did not show DIF when this was the only characteristic coded. It is not clear if the combination of VS3 with SE1 or VS3 alone is the primary cause of DIF. Five items were coded as SE1 and TM2 without VS3 and only one of these items showed DIF favoring students with mental retardation. There were no items coded as VS3 without SE1, so it was not possible to test the impact of verbal only items that did not require a social exchange. The items that require role play (SE2) and items that require the student to understand another's point of view (TM1) did not appear to result in DIF on ELA items. One of four items coded as SE2 and none of the four items coded as TM1 showed DIF. All of the other "Advantage MR" codes did not have sufficient items to test their merit.

Characteristics favoring MR on Mathematics

The accuracy of coding categories that were predicted to advantage MR was less clear on the mathematics items. Only two items were verbal only (VS3) and neither showed DIF, while 2 of the 4 items requiring a social exchange (SE1) showed DIF. The other social exchange codes which require symbolic play (SE2) or items that relate to social events (SE3) demonstrated DIF on 1 item each but these codes were only used a single time so results are far from conclusive. Twelve math items included the reuse of materials (PS1) and this characteristic did not appear to result in DIF (one item had DIF favoring MR, one item had DIF favoring AU, and 10 items showed no DIF). A similar pattern was found for the use of first or second person pronouns (TM2); the code was used for 16 items (2 had DIF favoring MR, 2 had DIF favoring AU, and 12 did not show DIF). The other coding categories were not used frequently enough to determine their merit.

Summary

Several of the hypothesized categories (either alone or in combination with other categories) were able to consistently predict DIF. Some of the categories that were hypothesized to result in DIF are directly related to the construct being assessed (e.g., the rote learning of phonics category is similar to the ELA standards of sight word reading and word analysis) and therefore should not be removed from the test. In other cases the characteristics hypothesized to result in DIF did not appear to be related to the constructs the test was intended to measure (e.g., items that required rote learning with longer attention and items that were verbally administered, required a social exchange, and used first or second person pronouns). Based on these findings, it would be advisable for test developers not to consider the use of test items that require a verbal response or those that require a social exchange unless it is determined that these are part of the construct being assessed. In addition a more detailed analysis of items that require rote memory and longer attention to task should be examined to determine if both of these characteristics are necessary to assess the intended construct. Although ELA items that focused on decoding unfamiliar words or reading sight words showed a strong relationship to DIF (favoring AU), these items are part of the construct being assessed and may need to be limited in number rather than eliminated entirely. Finally, the mathematics items that required a social exchange or social knowledge (SE1, SE2, or SE3) showed some DIF results; however the limited number of items did not make these findings conclusive.

Conclusions

Although several of the hypothesized coding categories resulted in accurate prediction of DIF, the study was limited to items from one testing program for students in one state. More research is needed to see if these hypotheses can be replicated across testing programs and populations and if the coding categories could be used to design assessments and test items that reduce the degree of DIF. Finally, previous research indicates that, in general, differences between students with autism and mental retardation are more pronounced in childhood than adolescence. In this study, there were more differences between Autism and Mental Retardation groups at the lower but not higher levels of the assessment. Future research is needed to examine the age by disability interaction effect leading to differential performance.

This study represents a first step in examining the performance of items designed for inclusion on alternate assessments for students with severe cognitive impairments. The results of the study may help to improve the nature and accessibility of the assessment and may also yield important information about the test performance of disability subgroups for parallel form construction.

References

American Psychiatric Association (2000). Diagnostic and Statistical Manual of Mental Disorders Fourth Edition Text Revision (DSM-IV-TR). Washington, DC: American Psychiatric Association.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-Aloud Accommodations: The Effect on Multiple Choice reading and Math Items* (Technical Report 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 2, 2004 from: http://education.umn.edu/NCEO/OnlinePubs/Technical31.htm

Bolt, S. E., & Bielinski, J. (2002, April). *The effects of the read-aloud accommodation on math test items.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Bolt, S. E., & Yesseldyke, J. E. (2006). Comparing DIF across math and reading/language arts test for students receiving a read-aloud accommodation. *Applied Measurement in Education, 19*(4), 329-355.

Cahalan-Laitusis, C., Cook, L., & Aicher, C. (April, 2004). *Examining Test Items for Students with Disabilities by Testing Accommodation on Assessments of English Language Arts*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*(2), 129-147.

Camilli, G. (1993). The case against item bias techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. (Research Report No. 91-47). Princeton, NJ: Educational Testing Service.

Individuals with Disabilities Education Act (IDEA), 20 U.S.C. S 1400 et seq. (1997).

Individuals with Disabilities Education Act (IDEA), 20 U.S.C. S 1400 et seq. (2004).

Koretz, D., & Hamilton, L. (2000). Assessment of Students with Disabilities in Kentucky: Inclusion, Student Performance, and Validity. *Educational and Policy Analysis, 22*, 255-272.

Maneckshana, B., Laitusis, C. C., & Monfils, L. (2006, April). *Comparisons of Disability Categories for an Alternate Assessment*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Muthen, B., Huang, L.-C., Jo, B., Khoo, S.-T., Goff, G. N., Novak, J. R., & Shih, J. C. (1995). Opportunity-to-Learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis, 17*(3), 371-403.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 U.S.C. § 1425 (2002).

No Child Left Behind Alternate Achievement Standards for Students with the Most Significant Cognitive Disabilities Non-Regulatory Guidance, Department Regulation 34 C.F.R. Part

200 (December 9, 2003). Available at www.ed.gov/legislation/FedRegister/finrule/2003-4/120903a.html

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: a review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*, 53-104.

Pitoniak, M. J., Cook, L. L., & Laitusis, C. C. (2006, April). *Using Differential Item Functioning Analyses to Investigate the Impact of Testing Accommodations on an English Language Arts Assessment*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 15, 2006, from: http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html

Shealy, R., & Stout, W. F. (1993a). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shealy, R., & Stout, W. F. (1993b). An item response model for test bias and differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S. G., Li, S., & Scarpati, S. (2003). The effects of test accommodations on test performance: A review of the literature. *Center for Educational Assessment Research Report No. 485*, Amherst, MA: School of Education, University of Massachusetts Amherst.

Thompson, S. J., Johnstone, C. J., Thurlow, M. L., & Altman, J. R. (2005). *2005 State special education outcomes: Steps forward in a decade of change*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 8, 2009 from: http://education.umn.edu/NCEO/OnlinePubs/2005StateReport.htm/

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Appendix

References to Support Coding Classifications

Baltaxe, C. (1977). Pragmatic deficits in the language of autistic adolescents. *Journal of Pediatric Psychology*, *2*, 176-180.

Dawson, G., & Adams, A. (1984). Imitation and social responsiveness in autistic children. *Journal of abnormal Child Psychology*, *12*, 209-226.

Frith, U., Happe, F., & Siddons, F. (1994). Autism and theory of mind in everyday life. *Social Development*, *3*(2), 108-124.

Henry, L. A., & MacLean, M. (2002). Working memory performance in children with and without intellectual disabilities. *American Journal on Mental Retardation, 107*, 421-432.

Jacobson, J. W., & Ackerman, L. J. (1990). Differences in adaptive functioning among people with autism or mental retardation. *Journal of Autism and Developmental Disabilities*, *20,* 205-219.

Klin, A., & Shepard, B. A. (1994). Psychological assessment of autistic children. *Child and Adolescent Psychiatry Clinics of North America, 3,* 131-148.

Koegel, L. K., Koegel, R. L., Hurley, C., & Frea, W. D. (1992). Improving social skills and disruptive behavior in children with autism through self-management. *Journal of Applied Behavior Analysis*, *25*, 341-353.

Koegel, R. L., Koegel, L. K., Frea, W. D., & Smith, A. E. (1995). Emerging interventions for children with autism. In Koegel & Koegel, (Eds.), *Teaching children with autism: Strategies for initiating positive interactions and improving learning opportunities* (pp. 1-15). Baltimore: Paul H. Brookes.

Leslie, A. M. (1992). Pretense, autism, and the theory-of-mind module. *Current Directions in Psychological Science*, *1*(1), 18-21.

Leslie, A. M., & Firth, U. (1987). Metarepresentation and autism: How not to lose one's marbles. *Cognition*, *27*, 291-294.

Losche, G. (1990). Sensorimotor and action development in autistic children from infancy to early childhood. *Journal of Childhood Psychology and Psychiatry, 31*, 749-761.

Maltz, A. (1981). Comparison of cognitive deficits among autistic and retarded children on the Arthur Adaptation of the Leiter International Performance Scales. *Journal of Autism and Developmental Disorders*, *11*, 413-426.

McArthur, D., & Adamson, L. B. (1996). Joint attention in preverbal children: Autism and developmental language disorders. *Journal of Autism and Developmental Disorders*, *26*, 481-496.

McCartney, J. R. (1987). Mentally retarded and nonretarded subjects' long-term recognition memory. *American Journal on Mental Retardation, 92,* 312-317.

McEvoy, R. E., Rogers, S. J., & Pennington, B. F. (1993). Executive function and social communication deficits in young autistic children. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 34*, 563-578.

National Research Council. (2001). *Educating children with autism*. Washington, DC: National Academy Press.

Numminen, H., Service, E., & Ruoppila, I. (2002). Working memory, intelligence, and knowledge base in adult persons with intellectual disability. *Research in Developmental Disabilities, 23,* 105-118.

Pearson, D. A., Yaffee, L. S., Loveland, K. A., & Lewis, K. R. (1996). Comparison of sustained and selective attention in children who have mental retardation with and without attention deficit hyperactivity disorder. *American Journal on Mental Retardation, 107,* 592-607.

Prior, M., & Ozonoff, S. (1998). Psychological factors in autism. In F. R. Volkmar (Ed.), *Autism and pervasive developmental disorders* (pp. 64-108). Cambridge, England: Cambridge University Press.

Sigman, M., & Ruskin, E. (1999). Continuity and change in the social competence of children with autism, Down syndrome, and developmental delays. *Monographs of the Society for Research in Child Development, 64*(1), 1-114.

Stone, W. L., Ousley, O., Yoder, P., Hogan, K., & Hepburn, S. (1997). Nonverbal communication in 2- and 3-year old children with autism. *Journal of Autism and Developmental Disabilities, 27*, 677-696.

Tager-Fluberg, H. (1996). Brief report: Current theory and research on language and communication in autism. *Journal of Autism and Developmental Disorders, 26*, 169-178.

Weatherby, A., & Prutting, C. (1984). Profiles of communicative and cognitive-social abilities in autistic children. *Journal of Speech and Hearing Research, 27*, 364-377.

Wetherby, A., Prizant, B., & Hutchinson, T. (1998). Communicative, social-affective, and symbolic profiles of young children with autism and pervasive developmental disorder *American Journal of Speech-Language Pathology, 7*, 79-91.

Wing, L., Gould, J., Yeates, R. R., & Brierley, L. M. (1977). Symbolic play in severely mentally retarded and in autistic children. *Journal of Child Psychology and Psychiatry, 18,* 167-178.