2009

The effects and perceived consequences of testing accommodations on math and science performance assessments

Stephen N. Elliott

Thomas R. Kratochwill

Brian McKevitt
University of Nebraska at Omaha, bmckevitt@unomaha.edu

Christine Kerres Malecki

Running head: EFFECTS OF TESTING ACCOMMODATIONS

The Effects and Perceived Consequences of Testing Accommodations on Math and Science Performance Assessments

Stephen N. Elliott

Learning Sciences Institute

Vanderbilt University

Thomas R. Kratochwill

University of Wisconsin-Madison

Brian C. McKevitt

University of Nebraska at Omaha

Christine Kerres Malecki

Northern Illinois University

Date Submitted: June 23, 2008
Date Resubmitted: February 23, 2009
Second Resubmission: July 30, 2009
Running Head: Effects of Testing Accommodations

Please address all correspondence concerning this manuscript to Brian C. McKevitt, Department of Psychology, University of Nebraska at Omaha, 6001 Dodge St., Omaha, NE 68182, (402) 554-2498 or bmckevitt@mail.unomaha.edu.

Footnote: This manuscript is based on the findings from a research project entitled An Experimental Analysis of the Effects of Testing Accommodations on the Scores of Students with Disabilities, funded by the Office of Special Education and Rehabilitative Services, United States Department of Education, Stephen N. Elliott and Thomas R. Kratochwill, Principal Investigators.

Abstract

The present study examined the effect accommodations have on test results of students with and without disabilities and documented experts' judgments about the appropriateness of testing accommodations. Test score data were collected from 218 fourth-grade students with and without disabilities on mathematics and science performance tasks and from eight testing experts who evaluated the fairness and validity of a sample of testing accommodations used with these students. Results indicated that, for most students with disabilities and some students without disabilities, packages of testing accommodations had a moderate to large effect on performance task scores. Expert reviewers rated most accommodations for a student with disabilities as being both valid and fair, and they gave accommodations listed on a student's IEP significantly higher validity and fairness ratings than accommodations that were not listed on the student's IEP. Interpretations of these data are provided and implications for practice and future research are discussed.

The Effects and Perceived Consequences of Testing Accommodations on Math and Science Performance Assessments

What do we know about the use and effects of testing accommodations for students with disabilities on performance assessments? While evidence continues to be gathered about the use and effects of testing accommodations on multiple-choice tests, little research has been done to examine how education and measurement experts perceive the validity and fairness of those accommodations when applied to specific students on performance assessments. The current study adds this unique contribution to the growing body of literature about testing accommodations.

Understanding the effects of testing accommodations on the test performance of students with disabilities is important to the overall validity of the test results. The No Child Left Behind Act of 2002 (NCLB) requires that *all* students participate in annual assessments to document student achievement and to hold schools accountable for student learning. Both NCLB and the Individuals with Disabilities Education Improvement Act of 2004 (IDEA 2004) also require students with disabilities meet academic proficiency standards set by each state. Therefore, meaningful participation in accountability assessment by students with disabilities is imperative for schools and states to meet the requirements of these laws. The use of accommodations facilitates this participation, and school personnel must make sure such accommodations provide fair and valid test results (Elliott, McKevitt, & Kettler, 2002).

Definition of Testing Accommodations

An *accommodation* is a change in an assessment that is intended to maintain or facilitate the measurement goals of the assessment so scores from the accommodated test would measure the same attributes as scores from the unaccommodated test (Elliott, Kratochwill, & Schulte,

1999; Elliott, Braden, & White, 2001). Any alteration to a test, however, may alter what is measured, and thus affect the validity of the inference made from the test results. Testing accommodations can include alterations to the presentation/response format, timing/scheduling, setting, and assistive technology. The term *modification* refers to alteration of test content. Some modifications will change what a test measures (i.e., change the construct) and consequently have a direct impact on the validity of inferences made from the test result (Elliott, Kettler, et al., in press). For example, reading a reading decoding test aloud to a student is a modification to the test that changes the construct; the reading decoding test actually becomes a listening test when read aloud to a student (McKevitt & Elliott, 2003).

In the present work, we were interested in researching the effect on student performances of *accommodations only* and on the perceived consequences of testing accommodations. One assumption of standardized tests is that they allow comparability across students because the test is administered in the same way, under the same conditions for all students. For some students, however, the validity of inferences made from test results may be questionable because aspects of a disability impeded performance on the test and the scores do not truly reflect the outcomes the test purports to measure. For this reason, tests are often altered in response to the characteristics of individuals' disabilities. Such alterations are said to "level the playing field" by removing irrelevant barriers to performance and allowing the person to demonstrate his or her "true" abilities. More accurately, testing accommodations are intended to *increase the validity of the inference made from a test score.*

At times, the characteristics of a student's disability might be directly related to the skills and knowledge intended to be measured. In these cases, establishing the validity of accommodations may be more challenging than when the accommodations remove only

construct-irrelevant barriers. A situation such as this highlights even more the importance of closely examining the student's characteristics, how those characteristics relate to the construct of the test, and what effect accommodations may have on the inferences made about the student's performance on the test.

Research on Testing Accommodations and Student Performance

Since the late 1990s, a substantial increase has occurred in research being conducted on the effects of testing accommodations. Tindal and Fuchs (1999) completed a comprehensive review of testing accommodations research and noted that most studies on accommodations have been large group designs, "making it impossible to predict the effect of any specific test change for an individual student" (p. 9). Therefore, more comprehensive, experimental studies that address the effects of accommodations on individual students were needed. The research most relevant to the current study are experimental studies in which researchers have compared the effects of testing accommodations on the performances of students with and without identified disabilities. Phillips (1994) argued that analyzing the effects of accommodations on students with and without disabilities is a useful way to demonstrate the effect an accommodation may have on the construct of a test. If an interaction between group (i.e., students with disabilities versus students without disabilities) and accommodation (i.e., accommodation present versus accommodation absent) exists, then the accommodation does not change the construct being measured. If, however, no interaction effect was present, then there would be no differential boost in scores for one group resulting from the accommodation, and scores could not be interpreted equally for the two groups. As McKevitt and Elliott (2003) pointed out, "with no interaction effect, the accommodation would either (a) benefit neither group or (b) benefit both

groups…[t]he desired effect is to have differential benefit, with the accommodations helping only those students who need them, thus creating an interaction effect" (p. 584.).

Schulte, Elliott, and Kratochwill (2001), for example, investigated the effects of testing accommodations on math test scores for students with and without disabilities. All students with a disability were paired with a student without a disability. Based on the IEP recommendations for accommodations for the student with a disability, each student in a pair were given those same accommodations. Two parallel forms of a standardized math test (TerraNova Multiple Assessment Battery; CTB/McGraw-Hill, 1997) were given to all participants, with one administration using the identified accommodations, and one administration utilizing no accommodations. The order of presentation (accommodated versus non-accommodated test administration) was randomized. Schulte et al. (2001) found that both students with disabilities and students without disabilities did significantly better on the accommodated test than on the non-accommodated test. Both groups of students had similar gains on performance-type items on the test, whereas students with disabilities had larger gains in their performance on the multiple-choice items on the test. Thus, based on Phillips' (1994) logic, because both groups improved their scores on the performance items, the accommodations then invalidated the resulting interpretations for these items. Because the multiple choice items demonstrated a differential boost, the accommodations could be interpreted as valid.

Elliott and Marquart (2004) conducted a similar study using the math TerraNova test focusing on only one accommodation, increased time to take a test. Unlike the previous study, no significant improvements were found in test scores for any students (with or without disabilities) when comparing the accommodated version of the test to the non-accommodated version of the test. The contrast in the Elliott and Marquart (2004) and Schulte et al. (2001)

studies indicates that just one accommodation (as common as extended time is) does not play a large enough role in isolation. Students need individualized sets of accommodations. This point about needing individualized sets, or packages, of accommodations is relevant to the current study; students in the current study received accommodation packages recommended by their teachers.

Conversely, Kettler et al. (2005) and Lang, Elliott, Bolt, and Kratochwill (2008) found improvements in the performance of students with and without disabilities when given accommodations recommended by teachers on standardized math and reading tests. In both studies, students with disabilities had overall greater score improvement with accommodations than students without disabilities with accommodations. However, the improvement in performance of students without disabilities when given individualized packages of accommodations as noted in these studies continues to give rise to doubt about the validity of accommodations in general. If accommodations function properly (i.e., remove a barrier to performance created by a disability), students without disabilities (i.e., those with no barriers) should not benefit significantly from their provision (Elliott et al., 2002).

The current study is a follow-up and extension of work done by Elliott, Kratochwill, and McKevitt (2001). These researchers investigated the performance of students with and without disabilities on math and science performance assessment tasks by comparing two conditions: the use of *individualized* accommodations noted on students' IEPs or as noted by teachers on the Assessment Accommodations Checklist (AAC; Elliott et al., 1999) or the use of no accommodations. Accommodations that were not on students' IEPs were used in some cases because teachers believed the IEPs were not comprehensive enough in their descriptions of appropriate accommodations. One group of students without disabilities also was given the tasks

with a "standard" accommodation package (i.e., extended time, help with directions and reading task text, and verbal encouragement). Results indicated that the teacher-recommended accommodations had the effect of improved test scores (medium to large effect sizes) for 75% of students with disabilities. Additionally, teacher-recommended accommodations resulted in improved scores (medium to large effect sizes) for 57% of students without disabilities. Fifty-five percent of students without disabilities with the "standard" accommodation package saw their scores significantly improved (medium to large effect sizes).

Clearly, a variety of findings have been observed across the studies reviewed. These findings may be due to the types of accommodations studied, the manner in which they were presented (e.g., isolated versus in packages), or the research designs used.. In addition, it should be noted that testing accommodations, when used properly, also should be used as instructional accommodations, so they are not new to the student in the testing situation. In the studies reviewed, it is unknown the degree to which the accommodations studied were already familiar to the students. Regardless, it is clear that research is still needed to test the effects of accommodations on the performance of students with disabilities, and studying the effect those accommodations have on students without disabilities is a useful way to do so.

Research on Teachers' and Other Experts' Perceptions of Testing Accommodations

There seems to be a common understanding that teachers' perceptions about the use of testing accommodations are an important factor to consider. To date, however, there have been few studies that have specifically examined the nature of those perceptions. Some interesting results, however, have been found in the empirical work that has been done in this area. First, Schulte, Elliott, and Kratochwill (2000) found that educators do not recommend using more testing accommodations for a student with a "severe" disability compared to a student with a

"mild" disability. Additionally, educators did not consider accommodations fairer for use with a student with a severe versus mild disability. Thus, educators' perceptions of accommodations for students of any type of disability seem to be very similar. Schulte et al. (2000) did find that educators recommended more accommodations and rated those accommodations as more fair and effective for performance-based assessments than for multiple-choice format assessments.

Another study compared the perceptions of general education teachers across grade levels (elementary, middle, and secondary) to examine trends (Jayanthi, Epstein, Polloway, & Bursuck, 1996). Jayanthi et al. investigated several aspects of teachers' attitudes regarding testing accommodations (ease of accommodating, who is responsible, etc.) and found that teachers perceived that some accommodations were more helpful than others and that elementary teachers found that providing some adaptations were easier than did middle and/or secondary level teachers. Regarding the fairness of testing adaptations being made only for students with disabilities, 67% of general education teachers rated it as "not fair" whereas 33% rated it as fair. The reason cited most often (78%) for it not being fair was the idea that *all* students need some adaptations.

Brackenreed (2004) investigated teachers' perceptions of the use of testing accommodations on a standardized test given in Ontario, Canada. Similar to the results from Jayanthi et al. (1996), Brackenreed found that teachers were concerned with fairness regarding the use of accommodations and the comparability of accommodated and unaccommodated test results. However, in their survey of teachers, McKevitt and Elliott (2003) found that teachers believed standardized testing was somewhat important and that accommodations improved the ability for students with disabilities to show what they know or can do on these tests. Similarly,

Lang et al. (2005) found that most teachers believed accommodations were fair and valid for students with disabilities.

Unfortunately, many teachers continue to have little knowledge about testing accommodations overall. Hollenbeck, Tindal, and Almond (1998) found that less than 55% of teachers in their survey knew what accommodations were allowable on the statewide assessment. Furthermore, teachers did not necessarily agree uniformly on which accommodations led to more or less valid results. Given that teachers still have concern over validity and fairness of accommodations, it would be useful to gain the perception of other measurement experts. Testing staff at CTB/McGraw-Hill (2002) developed a useful taxonomy for categorizing accommodations relative to the likelihood they would change the interpretation of the resulting test score if they were used. Category 1 accommodations (e.g., distraction-free space) should not impact score interpretation; Category 2 accommodations (e.g., extra time) may impact score interpretation; Category 3 accommodations (e.g., paraphrase directions) may affect not only score interpretation but also may change the construct that is being measured by the test. Using this taxonomy, Gibson, Haeberli, Glover, and Witter (2005) examined the frequency of IEP-listed accommodations recommended and used by teachers. They found the most frequently recommended accommodations were providing extra testing time and reading directions to students, both Category 2 accommodations. Therefore, the most commonly used accommodations may impact the validity and interpretation of the resulting score.

While this taxonomy is useful for categorizing the impact accommodations may have on test score interpretation, it relies solely on a description of the accommodation to judge its validity. It could be the case, however, that a testing expert who actually witnesses the provision of an accommodation may have different judgments about the accommodations validity and

fairness. In other words, how would individuals well-versed in psychometric properties of assessments rate the use of accommodations if they actually saw tests being delivered with accommodations? This question is addressed in the current study.

The Current Research Project

This study features data collected from 218 fourth-grade students participating in a testing accommodations research project over a period of 28 months. This sample included students with and without disabilities who received accommodations on challenging math and science performance assessment tasks (described below). In this design, students completed the assessments with and without accommodations, allowing for *intra*individual comparison of performance across accommodated and non-accommodated tasks. This alternating treatment design (ATD) also allows for *inter*individual comparisons of group performance. Effect sizes of accommodations are calculated for each student as the unit of analysis.

The main objectives for this study were to examine the effect accommodations have on test results of students with and without disabilities and document teachers' and measurement experts' judgments about the appropriateness of testing accommodations after testing is completed. As a result, we hypothesized that the IEPs of students with disabilities would contain at least one accommodation for standardized testing, and that several would contain a package of common accommodations such as extra time, tests read aloud, and quiet setting. We also hypothesized that teachers would select more accommodations for all students, regardless of disability status, for the performance assessment tasks. With regard to student performance, we believed that both students with and without disabilities would show score improvements on accommodated tasks, although greater gains would be shown by students with disabilities. Finally, we hypothesized measurement experts viewing the test administrations would believe

the accommodations were valid and fair for use on these tests, and that they would rate accommodations appearing on students' IEPs as more valid and fair than those that were not listed on the IEP.

Method

Participants

Participants in this study were 218 fourth-grade students. The students came from urban, suburban, and rural school districts in Wisconsin. Of the 218 students, 73 students had been diagnosed with a disability under Wisconsin eligibility criteria and were receiving special education. The majority of students with disabilities had a learning disability; because so few had other disabilities, all students with disabilities were combined into one group. Given that disability type should not be the basis for making accommodation decisions and has not been found to correspondent to specific accommodations (Elliott, 2007), this combining of students with disabilities into one group increased the power of our analyses. The remaining 145 students did not have disabilities and were included for comparison purposes. The educational histories of students without disabilities included in the study are unknown other than they were making acceptable progress in their 4th grade classrooms and nominated by teachers as "good students with no known disabilities." These students' parents also were interested in having them participate in the study. This study includes data from 100 students who also participated in the Elliott et al. (2001) study. Table 1 includes complete demographic data for these participants.

In addition, 10 adults were invited to participate in a video review of a student completing two tasks with accommodations to assess expert perceptions of using accommodations. Of the 10 persons invited, 8 completed the task. All eight expert reviewers had doctoral degrees in fields related to assessment and accommodations including school

psychology (n = 4), educational administration and special education (n = 3), and quantitative methods (n = 1). Five expert participants were female and three were male. Half of the participants were currently working as practitioners in education and half were professors at large, Western and Midwestern universities.

Materials

Two sets of performance assessment instruments, one for mathematics and a second for science, were developed by master teachers from the Wisconsin Student Assessment System project. These performance assessment instruments were designed to determine a student's competence in measurable learner outcomes relative to the state's content standards and to provide information to guide students and others in planning future education. These assessment tasks have been used for many years with a diverse sample of students and have known psychometric values (median alpha = .64, median r = .65). The tasks also have been found to be nonbiased and of nearly equal difficulty (using means and *sd*s) for students with and without disabilities (Braden, Elliott, & Kratochwill, 1998).

Each of the performance assessment instruments requires students to draw on their full range of knowledge of a content area to produce a response. In each instrument, students are required to apply their knowledge of concepts, skills, and procedures along with reasoning and higher-order thinking skills. Tasks were developed from the domains of mathematics defined by the National Council for Teachers of Mathematics (1989) and from the *National Science Education Standards* published by the National Research Council (1996). The tasks were aligned with Wisconsin's academic standards and were designed to challenge students to use problem solving and inquiry skills. As such, each task requires similar thinking and reasoning skills, despite differences in academic content.

The 4th-grade instruments require approximately four hours when completing all tasks. Scoring procedures for both the mathematics and science instruments are based on a 5-point descriptive set of performance criteria that range from 1 = Inadequate to 5 = Exemplary. Project assistants scored student responses using pre-established criteria. Each student's booklet was scored by two independent raters and adjudicated by a third rater if the first two ratings were not in agreement. Inter-rater reliability (kappa) for the first two raters was .65. The math and science tasks included in the booklets were:

Math Task 1: "The Race" asks students to analyze plans presented for a fair running race. Geometric knowledge and measurement skills are needed to complete the task.

Math Task 2: "Hot Dog" asks students to decide how many hot dogs and buns to buy for a picnic. The students need to read and interpret a table, use remainders in division, compute whole numbers, and estimate more than half of an odd number.

Math Task 3: "Latasha's Challenge" describes sisters playing with blocks to form cubes. One cube is missing in the pattern. Students are required to find and create the missing cube. They need to recognize a pattern and formulate a rule to extend the pattern.

Math Task 4: "Calculator Quest" describes a principal who wants to provide calculators for all the fourth graders while spending the least amount of money. This task requires students to (a) read and understand data in a table, and (b) compute amounts of money.

Science Task 1: "Creata Creature" asks students to evaluate the living conditions of a creature given certain biological and environmental characteristics. Students list their creature's adaptations, draw a picture of it, describe its habitat, and name the creature.

Science Task 2: “Temperature of Water” asks students to carry out an experiment using three cups of water and a thermometer. The students collect data and draw conclusions based on the information they collect and record.

Science Task 3: “Mobile” asks student to evaluate the balance of five mobile drawings. The students then create a new mobile drawing they believe would be balanced, test their mobile with manipulatives, and explain what they learned.

Science Task 4: “Nutrition" presents students with a food guide pyramid and asks questions about meals. The students need to read a diagram, classify items into groups, understand nutritional value, and compare and contrast.

The *Assessment Accommodations Checklist* (AAC; Elliott, Kratochwill, & Schulte, 1999) was used to document accommodations implemented. The AAC is a tool designed to assist IEP-team members with the selection of accommodations for use on tests by individual students. Teachers completed AACs for students with and without disabilities in this study to determine accommodations that would be used on the performance assessment tasks described above. Any accommodations listed on students’ IEPs were automatically recommended on the AAC. Teachers also had the option of choosing more accommodations on the AAC than those indicated on IEPs to ensure that students’ accommodation needs for the performance assessment tasks were adequately met.

Videotapes were used in the current study depicting one of the students with disabilities taking a math and a science assessment with accommodations. In addition, a modified version of the AAC was used to elicit experts’ perceptions of the accommodations they saw being used on the video. Specifically, the modification of the AAC resulted in a rating scale that involved indicating if an accommodation on the list was used in the video, rating the validity of that

accommodation for the observed student on the observed task (1 = not valid, 2 = validity is questionable, and 3 = valid), and rating the fairness of that accommodation for the observed student on the observed task (1 = not fair, 2 = fairness is questionable, and 3 = fair).

Procedure

All students with disabilities (*n* = 73; Condition 1) completed two math and two science tasks (randomly chosen) with accommodations recommended by teachers on IEPs or AACs provided to them. The remaining two math and two science tasks were completed independently without accommodations. The students without disabilities were randomly assigned to one of three test conditions: (1) no accommodations provided on any task (*n* = 46; Condition 2); (2) three common accommodations (considered a "standard package" of reading and paraphrasing directions, providing verbal encouragement, and giving extra time) provided on two math and two science tasks (*n* = 52; Condition 3); and (3) accommodations recommended by students' teachers provided on two math and two science tasks (*n* = 47; Condition 4). Students without disabilities in conditions 3 and 4 also completed the remaining two math and two science tasks independently without accommodations. The performance tasks were administered in four 1-hour sessions over the course of several weeks by teachers or project staff.

The purpose of including the "standard accommodation" condition was to gather data on the performance of students without disabilities using accommodations that we believed, based on data from a previous project, to be most frequently recommended by teachers. This performance can then be compared to the performance of those students who received accommodations based on individual need, to see if there is a difference in performance in students without disabilities when accommodations are individualized. The "standard package" of accommodations, consisting of reading and paraphrasing directions, providing verbal

encouragement, and giving extra time, was created based on data from a previous project (Elliott & Kratochwill, 1995-1998) that used the same performance assessment tasks. Data from that project revealed those three accommodations were the three most frequently provided to students on the tasks. The correspondence of these "standard" accommodations with those recommended on students' IEPs and recommended by teachers is discussed further in the results section. Furthermore, the "no accommodation" condition was included to ensure that sufficient numbers of students without disabilities completed all tasks without accommodations to enable meaningful comparisons for the effect size analyses.

Four of the students with disabilities who were assessed under Condition 1 (teacher-recommended accommodations) were randomly selected to have their assessment administrations videotaped. Thus, each videotape depicted one student taking a math and a science assessment (along with the test administrator for each task). One of those four videotapes was randomly selected and sent to 10 expert reviewers. The reviewers were directed to view the administration of the math and science task and to complete a modified AAC rating the validity and fairness of each accommodation they observed on the video. Thus, each participating expert reviewer (n=8) completed two modified AACs (one for their observation of the math task and one for their observation of the science task). Upon completion, the expert reviewers sent back the videotape and the completed modified AACs by mail. The expert reviewers were provided an honorarium for their participation.

Alternating Treatments Design and Data Analysis

An alternating Treatments Design (ATD; Hayes, Barlow, & Nelson-Gray, 1999) was used in which students were exposed to two different testing conditions in alternating fashion; i.e., students alternated between receiving and not receiving accommodations until all eight tasks

were completed. Starting condition (with or without accommodations) was randomly determined and task order was randomly assigned within each subject (math and science) to prevent order effects. As noted above, this design allows for both *intra*individual and *inter*group experimental comparisons without the need for baseline conditions. Each individual's performance was graphed, and effect sizes were calculated to examine the magnitude of the effects of accommodations on individual scores [see Busk & Serlin (1992) for a discussion of effect sizes in single-subject research].

Results

Information on Testing Accommodations Listed in Students' IEPs

Out of the 73 students with disabilities featured in this study, 38 had at least one accommodation recommended on their IEPs for use in standardized testing. The most frequently listed accommodation was extra test time, appearing on 24 IEPs, followed by reading directions to the student (22 IEPs), allowing the special education teacher to administer the test (22 IEPs), and rereading subtask directions to student (21 IEPs). Interestingly, only nine IEPs listed verbal encouragement as an accommodation. Nevertheless, these accommodations are generally consistent with those chosen for use in our "standard" test condition.

Testing Accommodations Used on Performance Assessment Tasks

Each student with a disability received an accommodation package based on the demands of the performance tasks as recommended by his or her teacher above and beyond those accommodations listed in the IEP. The accommodation packages typically consisted of 10 to 12 discrete accommodations. The most frequent accommodations used were verbal encouragement of effort (used with 60 students), read directions to student (60 students), simplify language in directions (55 students), reread subtask directions (54 students), have student restate directions to

the teacher in his/her own words (49 students), read test questions and content to student (46 students), and restate questions with more appropriate vocabulary (46 students). Extra time was necessary for 31 students with disabilities, but offered for almost all. Thus, these frequently-used accommodations are consistent with those used in our "standard" accommodation comparison group.

All 47 students without disabilities who received individualized accommodations had at least one accommodation recommended by their teacher. These students received an average of about five accommodations each during task administration. The most frequent accommodation provided for students without disabilities was verbal encouragement of effort (39 students), followed by read directions to student (27 students), student restate directions in his/her own words (26 students), and paraphrase or simplify language in directions (25 students). Extra testing time was used with only seven students without a disability.

Effects of Accommodations on Test Scores

Overall, when comparing non-accommodated test scores to accommodated test scores, testing accommodations had positive effects for 57 (78.1%) students with disabilities and 54 (54.5%) students without disabilities. "Small" or zero effects were found for 7 (9.6%) students with disabilities and 32 (32.3%) students without disabilities. Finally, negative effects were found for 9 (12.3%) students with disabilities and 13 (13.1%) students without disabilities. These data are summarized in Table 2. On average, the resulting individual effect size when comparing accommodated scores to non-accommodated scores was .88 (*sd* = .78) for students with disabilities, .44 (*sd* = .61) for students without disabilities who received the "standard" accommodations, and .45 (*sd* = .79) for students without disabilities who received teacher-recommended accommodations. Thus, there was virtually no increase in the scores of students

without disabilities who received individualized accommodations compared to those who received the "standard" package. Mean task scores and effect sizes reported for each test condition are presented in Table 3. Additionally, mean scores for students in all conditions are graphically reported in Figure 1. Note in Figure 1 that the performance of students with disabilities with accommodations is slightly lower than that of students without disabilities, while the performance of students with disabilities without accommodations is much below the average performance of other students.

Experts' Perceptions of Accommodations

Each of the eight reviewers watched the same student take a math and a science assessment using IEP recommended accommodations and rated the accommodations they saw being used. Validity and fairness ratings were compiled from the eight expert reviewers' modified AAC protocols. Table 4 displays descriptive data on all of the accommodations rated by at least one of the reviewers. A majority of the mean ratings on both the math and science tasks (on both validity and fairness) ranged from 2.50 to 3.00 (3 = valid/fair). The only ratings below 2.50 were on accommodations that were not listed on the AAC (i.e., the reviewers had to write in accommodations they saw being used that they could not match to an existing accommodation on the checklist).

It was hypothesized that accommodations used that were noted on the students' IEP would be rated as more valid and more fair than accommodations used that were not on the students' IEP. Table 5 presents mean rating scores collapsed by accommodation category (e.g., motivational accommodations, setting accommodations) and by IEP status (whether or not a group of accommodations were listed on the student's IEP). Because experts' ratings for the science and math task accommodations had few differences, the ratings for science and math

were combined to come up with one mean rating for validity and one mean rating for fairness in each accommodation category. Four overall scores were computed resulting in a total validity score for accommodations on the IEP, a total validity score for accommodations not on the IEP, a total fairness score for accommodations on the IEP, and a total fairness score for accommodations not on the IEP. A repeated measures ANOVA testing for a difference between the two validity scores was significant, Wilks' lambda = .357, F (1, 7) = 12.59, p = .009, with validity ratings on accommodations listed on the student's IEP being higher than ratings on accommodations not on the student's IEP. Similarly, an examination of the fairness ratings revealed a significant difference, Wilks' lambda = .382, F (1, 7) = 11.34, p = .012, with fairness ratings on accommodations listed on the student's IEP being higher than ratings on accommodations not on the student's IEP.

Discussion

This study contributes to the growing research base on testing accommodations and presents unique evidence using an Alternating Treatments Design to examine the effects of accommodations on students' scores on mathematics and science performance tasks. For the vast majority of students with disabilities, accommodations were provided in packages comprised of 10 to 12 discrete accommodations. It appears that these packages of testing accommodations have a moderate to large effect on challenging math and science performance assessment scores. For a small number of these students, effects were not positive. In general, individualized packages of accommodations similar to those used in actual test situations increased the scores of students with disabilities by ½ to 1 standard deviation or more. Some of these students were suspected of having academic difficulties (not assessed in the present study) and/or the tasks may have been too difficult for fourth-graders in general.

The findings of the current study support findings from previous studies that indicated accommodations are most typically provided in packages. The most common accommodations provided in this study, extra time, verbal encouragement, read directions aloud, and simplify language in directions, mirrored those observed by Elliott et al. (2001) and other studies reviewed. Findings pertaining to the effects of those accommodations also supported previous studies; over 70% of students with disabilities benefitted from the provision of accommodations, while around 55% of students without disabilities also benefitted. There was very little difference in performance of students without disabilities who received the standard package of accommodations versus those that were teacher recommended, with virtually identical mean effect sizes for the two groups. In practice, it may be the case that a standard package of accommodations could be sufficient for all students. This supposition needs further research involving students with disabilities rather than students without disabilities.

The reason for the increase in scores of students without disabilities continues to be unclear. It could be those students also had academic difficulties, or it could have to do with the difficulty of the tasks. Regardless, this research points out, as does research by Schulte et al. (2001), McKevitt and Elliott (2003), Kettler et al. (2005), and Lang et al. (2008), among others, that an increase in scores demonstrated by students without disabilities when provided accommodations may have an invalidating effect on test scores.

Unique to this study is the component examining experts' reviews of the provision of accommodations. Regarding expert reviewer ratings, most of the reviewers perceived the accommodations being used as being both "Valid" and "Fair," but slightly (but significantly) *more* valid and fairer when the accommodations were specified on the students' IEP. Without exception, the only accommodations rated as "not valid" or "not fair" by the expert reviewers

were those that they observed on the video that were unusual such as providing excessive prompts that may have resulted in guiding a student to the correct answer and providing samples or examples that were not part of the test protocol. These accommodations were seen more negatively on the math task than on the science task. In general, it appears that the experts viewed IEP-noted accommodations as fair and valid. In addition, the reviewers also perceived motivation (providing verbal encouragement), scheduling (extra time), setting (special education room), and format (large print) accommodations as being valid and fair even if they were not listed on the IEP.

The increase in scores of students without disabilities when provided accommodations found in this study and its predecessors raises interesting questions about the validity of the accommodations provided. Presumably, the accommodations should not affect the scores of students without accommodations because they would not have barriers (i.e., a disability) that interfere with test performance. These findings are worthy of future consideration regarding the validity of testing accommodations and the resulting test scores. However, by having measurement experts view the provision of accommodations, we have learned that even accommodations that may improve performance for students without disabilities are still viewed as fair and valid. Clearly, the discrepancy between the experts' evaluation of validity and the undesired score increases for students without disabilities is worthy of further investigation.

Having testing accommodations improve scores for students without disabilities can have a number of implications beyond just invalidating test scores. The improvement in score may mean that some students without disabilities have unidentified academic difficulties. The resulting score gain may serve to point out additional need for intervention or accommodation these students may have. The improvement in score may also point out that accommodations

could, and perhaps should, be used for more students to ensure that tests are truly measuring their knowledge and skills. Clearly there are policy implications that may stem from this latter point, but as research continues to point out the benefit accommodations have for some students with and without disabilities, it may be worthy of further consideration. For example, related research on accessible tests and item modifications is beginning to advance data-based evidence that many test items have extraneous information that when removed results in enhanced performances for many students (Elliott, Kettler, et al., in press). Finally, it may be the case that Phillips' (1994) interaction paradigm logic may not be the "true" test of the validity of scores with accommodated tests. Perhaps the improvement of students without disabilities when provided accommodations may not mean that the scores are invalid, but rather something might not be appropriate about the test being given, or even the instruction being provided.

There are several limitations to the current study worth noting. First, because accommodations were tested in packages, and not in isolation, it is difficult to know which accommodation(s) had the intended or unintended score effects. Furthermore, because students with and without disabilities may not have received the exact same accommodations, true comparisons between groups may be compromised. While individualized packages of accommodations are most commonly provided, future researchers should continue to study the effect of accommodations provided in isolation and provided equally to students with and without disabilities. Next, students' prior academic skills in math and science were not assessed. Performance on the tasks could have been impacted by previously learned skills. While the single participant nature of the research accounts for prior learning in many ways (i.e., each student is his or her own comparison), future research with more complete information about students' academic skills may be a useful addition to the study of the effects of accommodations.

Another concern with the present study was the rather large amount of variability observed in the effect sizes for the groups of students. This magnitude of variability has not been observed in other accommodation studies where multiple-choice items have been used, thus it is believed to be a function of the more difficult performance tasks. Future research, however, is left to test this explanation. Finally, eight measurement experts reviewed only one test administration in math and science. Conclusions based on more than one observation may be useful to truly understand the nature of the tasks and the accommodations to judge their fairness and validity appropriately. Future research should continue this interesting and informative perspective on the effects of accommodations.

As educators continue to try to address the mandates of NCLB and the meaningful participation of students with disabilities in a schools' achievement testing program, it is important to recognize the role of testing accommodations. This study focused on performance assessments in mathematics and science and adds to findings from previous studies that consistently have demonstrated that testing accommodations provided in "packages" to students can have moderate to large effects on their scores. This study also reminds us that despite the score increase, accommodations may be functioning to render the test results invalid or unfair, given that students without disabilities may also have improved scores from receiving accommodations. Nevertheless, educators must still make judgments about appropriate accommodations and should recommend valid and fair accommodations with good knowledge about the student being tested, the test being used, and state or district guidelines.

References

Brackenreed, D. (2004). Teacher's perceptions of the effects of testing accommodations. *Exceptionality Education Canada, 14,* 5-22.

Braden, J. P., Elliott, S. N. & Kratochwill, T. R. (1998, June). *The performance of students with and without exceptional educational needs on performance assessment and multiple choice achievement measures.* Paper presented at the annual meeting of the Council of Chief State School Officers Large Scale Assessment Conference, Colorado Springs, CO.

Busk, P. L., & Serlin, R. C. (1992). Meta-analyses for single-case research. In T. R. Kratochwill and J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Erlbaum.

CTB/McGraw-Hill. (1997). *TerraNova Multiple Assessment Battery.* Monterey, CA: Author.

Elliott, S.N. (2007). Selecting and using testing accommodations to facilitate meaningful participation of all students in state and district assessments. In L. Cook & C. Cahalan (Eds.). *Large-Scale Assessment and Accommodations: What Works?* (pp. 1-9). Princeton, NJ: ETS.

Elliott, S. N., Braden, J. P., & White, J. L. (2001). *Assessing one and all: Educational accountability for students with disabilities.* Alexandria, VA: Council for Exceptional Children.

Elliott, S.N., Kettler, R.J., Beddow, P.A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M., Bolt, D., & Roach, A.T. (in press). Using modified items to test students with and without persistent academic difficulties: Effects on groups and individualized students. *Exceptional Children*.

Elliott, S. N., & Kratochwill, T. R. (1995-1998). *Performance assessment and standardized testing for students with disabilities: Psychometric issues, accommodation procedures, and outcome analyses.* (U.S. Department of Education Grant No. 84.023F). University of Wisconsin-Madison: Wisconsin Center for Education Research.

Elliott, S. N., & Kratochwill, T. R. (1998-2000). *Experimental analysis of the effects of testing accommodations on the scores of students with disabilities.* (U.S. Department of Education Grant No. 84.023). University of Wisconsin-Madison: Wisconsin Center for Education Research.

Elliott, S. N., Kratochwill, T. R., & McKevitt, B. C. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology, 3*9, 3-24.

Elliott, S. N., Kratochwill, T. R., & Schulte, A. G. (1999). *The assessment accommodations checklist.* Monterey, CA: CTB/McGraw Hill.

Elliott, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children, 70*(3), 349-367.

Elliott, S. N., McKevitt, B. C., & Kettler, R. (2002). Testing accommodations research and decision-making: The case of "good" scores being highly valued but difficult to achieve for all students. *Measurement & Evaluation in Counseling & Development, 35*, 156-166.

Gibson, D., Haeberli, F. B., Glover, T. A., & Witter, E. A. (2005). Use of recommended and provided testing accommodations. *Assessment for Effective Instruction, 31(1),* 19-36.

Hayes, S. C., Barlow, D. H., & Nelson-Gray, R. O. (1999). *The scientist practitioner: Research and accountability in the age of managed care.* Boston: Allyn & Bacon.

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

Jayanthi, M., Epstein, M.H., Polloway, E.A., Bursuck, W.D. (1996). A national survey of general education teachers' perceptions of testing adaptations. *The Journal of Special Education, 30,* 99-115.

Kettler, R. J., Niebling, B. C., Mroch, A. A., Feldman, E. S., Newell, M. L., Elliott, S. N., Kratochwill, T. R., & Bolt, D. M. (2005). Effects of testing accommodations on math and reading scores: An experimental analysis of the performance of students with and without disabilities. *Assessment for Effective Intervention, 31,* 37-48.

Lang, S. C., Elliott, S. N., Bolt, D. M., & Kratochwill, T. R. (2008). The effects of testing accommodations on students' performances and reactions to testing. *School Psychology Quarterly, 23,* 107-124.

Lang, S. C., Kumke, P. J., Ray, C. E., Cowell, E. L., Elliott, S. N., Kratochwill, T. R., & Bolt, D. M. (2005). Consequences of using testing accommodations: Student, teacher, and parent perceptions of and reactions to testing accommodations. *Assessment for Effective Intervention, 31,* 49-62.

McKevitt, B. C., & Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended testing accommodations on a reading achievement test. *School Psychology Review, 32,* 583-600.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Research Council (1996). *National science education standards.* Washington, DC: National Academy Press.

Phillips, S. E. (1994). High stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7,* 93-120.

Schulte, A.A.G., Elliott, S.N., Kratochwill, T.R. (2000). Educators' perceptions and documentation of testing accommodations for students with disabilities. *Special Services in the Schools, 16,* 35-56.

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.

Tindal, G., & Fuchs, L. (1999). *A summary of research on test changes: An empirical basis for defining testing accommodations*. Mid-South Regional Resource Center, Lexington, KY.

Figure Caption

Figure 1. Mean performance scores of students with disabilities with and without accommodations and students without disabilities without accommodations.

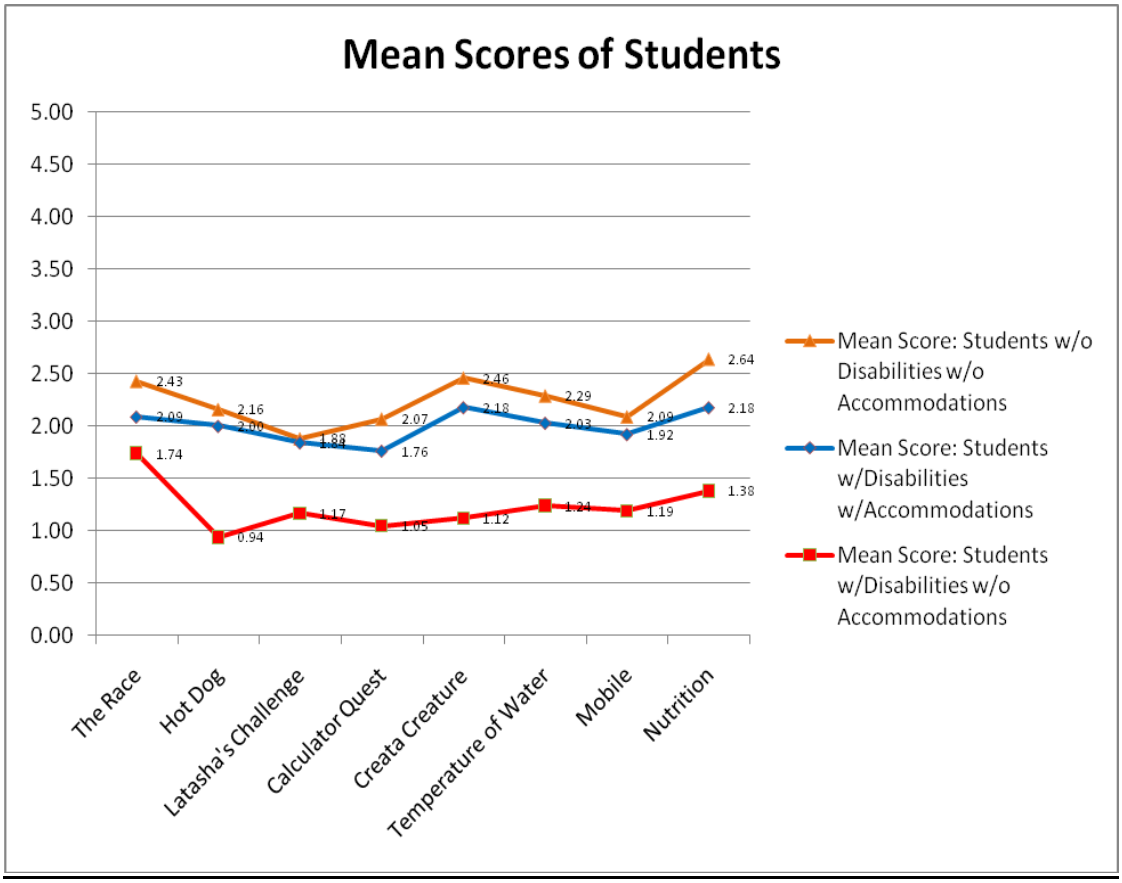
Mean Scores of Students
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
2.43
2.16
2.07
2.46
2.29
2.09
2.64
2.09
2.00
1.88
1.76
2.18
2.03
1.92
2.18
1.74
0.94
1.17
1.05
1.12
1.24
1.19
1.38
The Race
Hot Dog
Latasha's Challenge
Calculator Quest
Creata Creature
Temperature of Water
Mobile
Nutrition
Mean Score: Students w/o Disabilities w/o Accommodations
Mean Score: Students w/Disabilities w/Accommodations
Mean Score: Students w/Disabilities w/o Accommodations

Table 1

Description of Student Participants

	Number in Sample (n = 218)	Percentage of Sample
Gender		
Female	102	46.8
Male	116	53.2
Ethnicity		
African-American	15	6.9
Multi-Ethnic	4	1.8
White	196	89.9
Hispanic	2	0.9
Native American	1	0.5
Disability		
None	146	67.0
Learning Disability	43	19.7
Emotional Disturbance	7	3.2
Cognitive Disability	3	1.4
Speech/Language	9	4.1
Other Health Impairment	1	0.5
Autism	2	0.9
Learning Disability & Speech/Language	3	1.4
Cognitive Disability & Emotional Disturbance	1	0.5
Cognitive Disability & Hearing Impairment	1	0.5
Other Health Impairment & Speech/Language	2	0.9
Condition		
1. Students with Disabilities: Teacher-Recommended and/or IEP accommodations	73	33.5
2. Students without Disabilities: No Accommodations	46	21.1
3. Students without Disabilities: Standard Accommodations Package	52	23.9
4. Students without Disabilities: Teacher-Recommended Accommodations	47	21.6

Table 2

Accommodation Effect Sizes

Effect Size*	Students with Disabilities (n = 73)	Students without Disabilities: Standard Accommodations (n = 52)	Students without Disabilities: Teacher-recommended Accommodations (n = 47)
Large (greater than .80)	46 (63%)	17 (32.6%)	16 (34%)
Medium (.40 - .80)	11 (15.1%)	10 (19.3%)	11 (23.4%)
Small (less than .40 but greater than 0)	2 (2.8%)	8 (15.4%)	3 (6.4%)
Zero	5 (6.8%)	11 (21.2%)	10 (21.3%)
Negative	9 (12.3%)	6 (11.5%)	7 (14.9%)

*Note: The effect size designation as large, medium, and small is somewhat arbitrary and does not follow criteria advanced for group designs since these data are based on single-case design.

Table 3
Mean Task Scores and Effect Sizes Per Condition

	Students with Disabilities (n = 73)		Students without Disabilities (n = 145)				
			No Accommodations (n = 46)	Standard Accommodation Package (n = 52)		Teacher-Recommended Accommodations (n = 47)	
Task	Without Accommodations	With Accommodations	Without Accommodations	Without Accommodations	With Accommodations	Without Accommodations	With Accommodations
Math							
The Race	1.74 (.89)	2.09 (.70)	2.72 (.75)	2.63 (.56)	2.68 (.65)	2.43 (.51)	2.67 (.56)
Hot Dog	0.94 (.69)	2.00 (.81)	2.39 (.61)	2.13 (.56)	2.40 (.93)	2.16 (.90)	2.75 (.75)
Latasha's Challenge	1.17 (.62)	1.84 (.86)	1.80 (.62)	1.84 (.62)	2.07 (.73)	1.88 (.53)	2.27 (.70)
Calculator Quest	1.05 (.83)	1.76 (.61)	2.24 (.71)	2.04 (.66)	2.54 (.65)	2.07 (.62)	2.50 (.76)
Science							
Creata Creature	1.12 (.91)	2.18 (.68)	2.59 (.78)	2.46 (.72)	2.79 (.79)	2.46 (.81)	2.71 (.78)
Temperature of Water	1.24 (.86)	2.03 (.93)	2.09 (.76)	2.13 (.83)	2.65 (.81)	2.29 (1.06)	2.42 (.76)
Mobile	1.19 (.86)	1.92 (.68)	2.22 (.55)	2.19 (.63)	2.38 (.57)	2.09 (.61)	2.28 (.61)
Nutrition	1.38 (1.02)	2.18 (.60)	2.85 (.63)	2.55 (.67)	2.77 (.63)	2.64 (.70)	2.68 (.57)
Mean Individual Effect Size	**0.88 (.78)**			**0.44 (.61)**		**0.45 (.79)**	

Note. Standard deviation in parentheses.

Table 4

Expert Perceptions of the Validity and Fairness of Testing Accommodations

			Math Task						Science Task					
			Valid			Fair			Valid			Fair		
On IEP?	Category	Accommodation	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
no	Motivation	Verbal encouragement of effort	3.00	.00	8	3.00	.00	8	3.00	.00	5	3.00	.00	5
no	Motivation	Encourage to stay on task	--	--	-	--	--	-	3.00	--	1	3.00	--	1
no	Motivation	Encouragement when slow starting	3.00	--	1	3.00	--	1	--	--	-	--	--	-
yes	Scheduling	Provide extra test time	3.00	.00	4	3.00	.00	4	3.00	.00	3	3.00	.00	3
no	Setting	Distraction free space	3.00	.00	8	3.00	.00	8	3.00	.00	7	3.00	.00	7
yes	Setting	Special ed classroom	3.00	.00	2	3.00	.00	2	3.00	.00	2	3.00	.00	2
no	Setting	Individual test administration	3.00	.00	8	2.87	.35	8	3.00	.00	8	3.00	.00	8
yes	Directions	Read directions	3.00	.00	8	3.00	.00	8	2.87	.35	8	2.87	.35	8
yes	Directions	Read subtask directions	3.00	.00	6	3.00	.00	6	2.80	.45	5	2.80	.45	5
no	Directions	Simplify language in directions	2.75	.46	8	2.63	.52	8	2.87	.35	8	2.87	.35	8
no	Directions	Clarify questions by asking	3.00	.00	2	3.00	.00	2	3.00	.00	3	3.00	.00	3
no	Directions	Have student reread/restate directions	3.00	--	1	3.00	--	1	3.00	--	1	3.00	--	1

			Math Task						Science Task					
			Valid			Fair			Valid			Fair		
On IEP?	Category	Accommodation	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*
yes	Assessment	Read questions and content	3.00	.00	7	3.00	.00	7	2.86	.38	7	2.86	.38	7
no	Assessment	Restate question w/appropriate vocab.	2.63	.52	8	2.63	.52	8	3.00	.00	5	3.00	.00	5
no	Assessment	Turn pages for student	3.00	.00	3	3.00	.00	3	--	--	-	--	--	-
no	Assessment	Assist in tracking items by pointing	3.00	--	1	3.00	--	1	3.00	.00	2	3.00	.00	2
no	Assessment	Have teacher sit near student	2.86	.38	7	2.86	.38	7	3.00	.00	7	3.00	.00	7
no	Equipment	Manipulatives	2.57	.53	7	2.57	.53	7	3.00	.00	2	3.00	.00	2
no	Format	Large-print answer document	3.00	--	1	3.00	--	1	--	--	-	--	--	-
no	Format	Mark responses in test booklet	3.00	.00	2	3.00	.00	2	3.00	--	1	3.00	--	1
no	Other	Prompts/redirection guiding to answer	1.57	.53	7	1.71	.76	7	2.67	.58	3	2.33	.58	3
no	Other	Provided sample/example not on test	1.00	.00	3	1.33	.58	3	2.00	.00	2	2.50	.71	2

Note. 1 = Not Valid/Not Fair, 2 = Validity/Fairness is Questionable, 3 = Valid/Fair. “On IEP” refers to whether or not the accommodation was actually listed on the student’s IEP.

Table 5

Accommodation Category Mean Ratings and Total Validity and Fairness Ratings

Category	On IEP			Not on IEP		
	M	*SD*	*N*	*M*	*SD*	*N*
Motivation						
Validity	n/a	n/a	n/a	3.00	.00	8
Fairness	n/a	n/a	n/a	3.00	.00	8
Scheduling						
Validity	3.00	.00	4	n/a	n/a	n/a
Fairness	3.00	.00	4	n/a	n/a	n/a
Setting						
Validity	3.00	.00	2	3.00	.00	8
Fairness	3.00	.00	2	2.96	.12	8
Directions						
Validity	2.94	.18	8	2.85	.35	8
Fairness	2.94	.18	8	2.79	.36	8
Assessment						
Validity	2.93	.19	7	2.88	.14	8
Fairness	2.93	.19	7	2.86	.16	8
Equipment						
Validity	n/a	n/a	n/a	2.71	.39	7
Fairness	n/a	n/a	n/a	2.64	.48	7
Format						
Validity	n/a	n/a	n/a	3.00	.00	2
Fairness	n/a	n/a	n/a	3.00	.00	2
Other						
Validity	n/a	n/a	n/a	1.75	.56	7
Fairness	n/a	n/a	n/a	1.82	.69	7
Total						
Validity	**2.97**	**.09**	**8**	**2.73**	**.12**	**8**
Fairness	**2.97**	**.09**	**8**	**2.72**	**.14**	**8**

Note. Scores reflect both Math and Science task accommodation ratings.