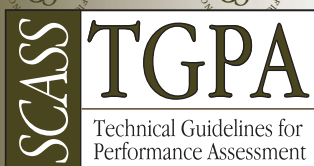




State Collaborative on Assessment
and Student Standards

Accommodating Mathematics Testing Using a Videotaped, Read-Aloud Administration

Gerald Tindal



Technical Guidelines for
Performance Assessment

A Publication of the Council of Chief State School Officers



The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its structure of standing committees and special task forces, the Council responds to a broad range of concerns about education and provides leadership on major education issues.

Because the Council represents each state's chief education administrator, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. CCSSO forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

The State Education Assessment Center was established through a resolution by the membership of CCSSO in 1984. This report is sponsored by the Assessment Center's State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA) consortium. The SCASS TGPA works with researchers to design and implement practical and timely research on large-scale performance assessment. This research provides information useful in designing state assessment and accountability programs so that they yield results that can be used to improve student learning.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

G. Thomas Houlihan
Executive Director

Wayne N. Martin
Director
State Education Assessment Center

John F. Olson
Director of Assessments
State Education Assessment Center

Phoebe C. Winter
Project Director, Technical Guidelines for Performance Assessment
State Collaborative on Assessment and Student Standards



Accommodating Mathematics Testing Using a Videotaped, Read-Aloud Administration

Gerald Tindal
University of Oregon

With the TGPA SCASS Study Group
and
the University of Oregon Center on Behavioral Research and Teaching

This report was prepared for submission under contract with the Council of Chief State School Officers. The preparation of the report was funded by OERI Grant No. R279A50006. The views and opinions expressed in this report are not necessarily those of the United States Department of Education, the Council of Chief State School Officers, the State Collaborative on Assessment and Student Standards, or the states participating in the study.

Council of Chief State School Officers
One Massachusetts Avenue, N.W.
Washington, DC 20001

February, 2002

Acknowledgements

CCSSO INDIVIDUALS AND TGPA SCASS STUDY GROUP MEMBERS

Pat Almond	Oregon Department of Education
Sue Bechard	Measured Progress (formerly with Colorado Department of Education)
Peter Behuniak	Connecticut Department of Education
Fen Chou	Louisiana Department of Education
Jan Hasbrouck	Texas A&M, College Station
Ellen Hedlund	Rhode Island Department of Education
Ellie Sanford	MetaMetrics (formerly with North Carolina Department of Public Instruction)
Alan Sheinker	CTB/ McGraw-Hill (formerly with Wyoming Department of Education)
Martha Thurlow	National Center on Educational Outcomes
Gloria Turner	Alabama Department of Education, Study Group Chair
Rebecca Walk	Wyoming Department of Education
Phoebe Winter	Council of Chief State School Officers
Liru Zhang	Delaware Department of Education

UNIVERSITY OF OREGON BEHAVIORAL RESEARCH AND TEACHING

Aaron Glasgow
William Heath
Raina Megert
Sarah McCully
Robert Helwig, Ph.D.
Keith Hollenbeck, Ph.D.
Dae-Sik Lee, Ph.D.

Contents

ABSTRACT	1
ACCOMMODATING MATHEMATICS TESTING USING A VIDEOTAPED, READ-ALoud ADMINISTRATION	3
METHOD	7
Participants	7
Table 1. Demographic Characteristics Of Participants By Grades (Elementary And Middle) – Count (%)*	8
Table 2. Educational Status Of Participants by Grades (Elementary and Middle) Count (%)*	9
Measures	9
Table 3. Descriptive Statistics For Proportion Correct With Coefficient Alpha For Forms In Elementary And Middle Level Test.	9
Procedures	11
Data Analysis	12
Order and Form Effects	13
Table 4. Descriptive Statistics for Order from the Standard Administration for Elementary School Students	13
Table 5. Mean Differences for Order from the Standard Administration for Elementary School Students	13
Table 6. Descriptive Statistics for Order from the Video Administration for Elementary School Students	13
Table 7. Mean Differences for Order from the Video Administration for Elementary School Students	14
Table 8. Descriptive Statistics for Order from the Standard Administration for Middle School Students	14
Table 9. Mean Differences for Order from the Standard Administration for Middle School Students	14
Table 10. Descriptive Statistics for Order from the Video Administration for Middle School Students	14
Table 11. Mean Differences for Order from the Video Administration for Middle School Students	14
Table 12. Descriptive Statistics for Form from the Standard Administration for Elementary School Students	15
Table 13. Mean Differences for Form from the Standard Administration for Elementary School Students	15
Table 14. Descriptive Statistics for Form from the Video Administration for Elementary School Students	15
Table 15. Mean Differences for Form from the Video Administration for Elementary School Students	15
Table 16. Descriptive Statistics for Form from the Standard Administration for Middle School Students	15
Table 17. Mean Differences for Form from the Standard Administration for Middle School Students	16
Table 18. Descriptive Statistics for Form from the Video Administration for Middle School Students	16
Table 19. Mean Differences for Form from the Video Administration for Middle School Students	16
Analysis of Administration Method	16
Elementary Level	16
Table 20. ANOVA Table for Standard and Video Administration by Elementary Student Classification and Low Reading	16
Table 21. Means Table for Standard and Video Administration by Elementary Student Classification and Low Reading	16
Table 22. ANOVA Table for Standard and Video Administration by Elementary Student Classification	17
Table 23. Means Table for Standard and Video Administration by Elementary Student Classification	17
Table 24. Frequency Distribution of Difference between Video-Standard Administration by Elementary Student Classification	18
Middle School Level	19
Table 25. ANOVA Table for Standard and Video Administration by Middle Student Classification and Low Reading	19
Table 26. Means Table for Standard and Video Administration by Middle Student Classification and Low Reading	19
Table 27. ANOVA Table for Standard and Video Administration by Middle Student Classification	19
Table 28. Means Table for Standard and Video Administration by Middle Student Classification	19
Table 29. Frequency Distribution of Difference between Video-Standard Administration by Middle Student Classification	20
Relationship with Criterion Measures	21
Table 30. Descriptive Statistics of Three Criterion Measures Used as Marker Variables in Reading and Math Skills for Elementary School Students (No LEP or 504): z-scores and raw scores, respectively	21
Table 31. Descriptive Statistics of Three Criterion Measures Used as Marker Variables in Reading and Math Skills for Middle School Students (No LEP or 504): z-scores and raw scores, respectively	22

<i>Table 32. Correlations Among Criterion Measures and Multiple-Choice Test (Both Administration Conditions & No LEP or 504) Using Grade Level Standard Scores for Elementary School Students</i>	<i>22</i>
<i>Table 33. Correlations Among Criterion Measures and Multiple-Choice Test (Both Administration Conditions & No LEP or 504) Using Grade Level Standard Scores for Middle School Students</i>	<i>23</i>
<i>Table 34. Frequency of Rating Values on Open-Ended Math Performance Task for Elementary School Students (No LEP or 504)</i>	<i>23</i>
<i>Table 35. Frequency of Rating Values on Open-Ended Math Performance Task for Middle School Students (No LEP or 504)</i>	<i>23</i>
<i>Table 36. Correlation for Elementary Students between Open-Ended Math Performance Task and Standard and Video Administration of the Multiple-Choice Test using Spearman Rank (No LEP or 504 Students).....</i>	<i>24</i>
<i>Table 37. Correlation for Middle Students between Open-Ended Math Performance Task and Standard and Video Administration of the Multiple-Choice Test using Spearman Rank (No LEP or 504 Students).</i>	<i>25</i>
DISCUSSION	27
Interpretations	27
Reflections.....	29
REFERENCES	33

With the mandates of the Individuals with Disabilities Education Act of 1997, students with disabilities must participate in large-scale assessment. When necessary, accommodations must be used to ensure that their participation is appropriate. The purpose of an accommodation is to ensure access, remove barriers and impediments, and provide a measure of behavior that leads to valid inferences. Given the importance of large-scale assessment and the current emphasis on including all students, research on test accommodations needs to establish which accommodations are or are not appropriate. In this study, a mathematics test was read aloud using a videotaped presentation in which problems were presented singly, in a paced format, and with visual prompting of the answer choices. Students' performance on this accommodated presentation was compared to their performance with a standard administration in which students read the problems themselves with several multiple-choice items presented on a page and no pacing was used. Students participated in both administrations, requiring two different forms of a test to be given in counterbalanced order. For students in fourth and fifth grades, performance was higher with the videotaped presentation than with the standard administration for both general education students rated low in reading and those with IEPs in reading. In seventh and eighth grades, no such improvements appeared with the videotaped administration for either group of students. These results are interpreted not only in the context of making valid inferences about students' performance, but also in terms of the manner we measure mathematics skills.

Accommodating Mathematics Testing Using a Videotaped, Read-Aloud Administration

In a recent publication from the Mid-South Regional Resource Center, Tindal and Fuchs (1999) review the research on test accommodations that has been accumulating over the past two decades. This publication is designed to establish an empirical basis for the recommendation for and implementation of test accommodations, as “there does not currently exist a set of guidelines about acceptable accommodations that is based on comprehensive empirical research. This is because we do not have a comprehensive set of research on testing accommodations...and that there currently exists little consistency in assessment policy” (Thurlow, Ysseldyke, & Silverstein, 1993, p. 3).

The National Center on Educational Outcomes (NCEO) at the University of Minnesota has been tracking state policies and practices on statewide testing for nearly a decade, including what kinds of changes are permitted and what kinds are disallowed. Assessment policies range from allowing no changes to be made in the tests to permitting specific changes for some students. Typically, when the changes are significant and the construct being measured is likewise changed, the term modification is used. In contrast, if the changes are minimal and the construct being measured is not altered, the term “accommodation” is used. Like the issue of inclusion, this distinction in testing and measurement practices is rife with controversy. Accommodations and modifications have been grouped into the following four categories (Ysseldyke, Thurlow, McGrew, & Shriner, 1994):

- Presentation, in which the stimuli (materials) presented to students are changed.
- Response, in which students are allowed to use a different manner of responding.
- Setting, in which the context of where tests are administered and who administers the test is varied.
- Timing and scheduling, in which changes are made in how long a student has to take the test and in how many sessions are administered.

NCEO describes three types of students with disabilities in considering accommodations of testing practices and procedures. Some students can take large-scale assessments with no accommodations. Other students can be included but require accommodations and adaptations. Finally, a small group of students need to take a different assessment because their curriculum is different from that tested. They suggest, however, that about 85% of the students with disabilities comprise the first two groups.

In this study, we describe a change in testing students’ mathematics proficiency that was designed to be an accommodation, not a modification. We used a videotaped read-aloud of mathematics problems and answer options for two reasons. First, reading should not interfere with the primary construct of mathematics. Second, videotape would allow other accommodations to be used (e.g., in timing, scheduling, or setting). For example, with

videotape, a number of setting accommodations could be implemented (administer the test in a separate location individually, with a small group, or with minimal distractions). Likewise, timing/scheduling accommodations also could be considered (allow frequent breaks during testing, administer the test in several sessions, or change the time of day). From a practical point of view, such accommodations would be very difficult to provide for many students without videotape.

In the accommodation investigated in this study, a multiple-choice mathematics test was read out loud to students using a videotaped presentation. This study follows a line of previous research: Six other studies have been done in which students had the test read to them. Two of these published reports, on the same study, were from two different authors and resulted in different conclusions. Koretz (1997) reported on oral reading (along with rephrasing, cueing, and dictation) of mathematics and science tests for fourth and eighth grade students taking the Kentucky Essential Skills Test. He concluded that the test was biased given that students with moderate cognitive and learning disabilities who received the accommodation scored near the mean of students without disabilities and who did not receive the accommodation. In contrast, Trimble (1998) reported that only 4 significant differences appeared from the 104 comparisons that were made comparing students' performance with and without the accommodations. In this research, the reading aloud accommodation was part of a package in which other accommodations also were used (dictation, rephrasing, and cueing) and statistical estimates only were available for documenting its unique effect.

A study by Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) reported that fourth grade students with learning disabilities and Individualized Education Programs (IEPs) in reading improved significantly on their mathematics performance when the test was read aloud to them. In fact, these authors reported a significant interaction, in which no such performance improvement was reported for students without disabilities (and ranked as the lowest 10 in the classroom) who received the same accommodation. The difference between general and special education students was significant when students read the multiple-choice test silently and no such difference was found between the two groups when the teachers read the problems and options. This interaction is an important component of this study: Students in special education had IEPs in reading and those in general education had been ranked in the lowest group of 10 in their classroom. Therefore, the findings provide initial support for accommodations as providing access.

A follow-up study by Helwig, Tedesco, Heath, Tindal, and Almond (1999) with a sixth-grade student population revealed no significant effects. Overall performance on a multiple-choice mathematics test was the same whether or not it was read out loud to students. This study, however, presented the read-aloud administration with videotape. Students' reading and mathematics ability were measured by a pre-mathematics and reading test. Only students who scored average or above on the mathematics ability test were included in the data analysis. Students were divided into three groups (high, medium and low reading ability) based on their Oral Reading Fluency (ORF) scores. For data analysis, the percentage of students answering each problem correctly was calculated. Because each problem appeared on both versions of the test, it was possible to calculate the difference in percentage correct between the standard administration and the video version. The number of multi-syllable (MS) words in each problem was calculated. A total of 24 problems had relatively few MS words and 19 problems had relatively many MS words. They found that three of the four groups performed better on the standard version for low MS questions. All groups performed better on the video version of the test on high MS questions. The magnitude of the difference between high and low MS questions was only significant for low readers. These findings suggest that, with certain types of mathematics problems, a read-aloud accommodation can be effective for

low ability readers. However, average and above ability readers may find a video administration distracting. These findings contribute to our understanding of the constructs being measured in our tests of basic skills.

Fuchs, Fuchs, Eaton, Hamlett, and Karns (1998) also reported significant improvement when elementary students had a mathematics curriculum-based measure read to them (along with extended time to complete the test and the use of a calculator). This same accommodation, however, was not effective on a traditional achievement test of concepts and applications. They also reported that in comparing recommendations to provide an accommodation, teachers tend to over recommend them, and often the impact on student performance is negligible.

Finally, Weston (1999) reported positive differential results when fourth grade students took a mathematics test that was read to them. He found that on multiple-choice mathematics problems, whether calculation oriented or based on word story problems, performance improved when the problems were read out loud. Although significant main effects were found for both groups of students and both forms of the test, an interaction also was found rendering the main effect unimportant: A larger effect from the accommodation was found for students with disabilities. Weston noted, "Much of the effect for learning disabled students occurs at lower reading levels where regular education students are not well represented in the study" (p. 9). Finally both types of mathematics problems, the calculation as well as the word problems, showed an effect from the reading accommodation.

Although this research helps establish the validity of inferences from administering mathematics multiple-choice tests by reading them to students, we also added three other accommodations in the current study. We attempted to justify the changes based on the research literature and otherwise made the decisions to change the test so that the primary intervention (read aloud of mathematics problems) had the best probability of being implemented.

We began the study to compare two different methods of administration for a mathematics multiple-choice test: a videotaped read-aloud versus a standard administration. As part of the experimental design, we allocated our resources to avoid threats to validity. Students were crossed with the treatment (method of test administration), alternate forms of the test were used so that each form was administered in both the videotaped and the standard method, and the order of administration of each of these four combinations was counterbalanced. Teachers were systematically trained in the administration of the test in both conditions and all student protocols were machine scored. In the end, this design allowed us to carefully analyze the results with four factors: order, form, grade level (elementary versus middle school as well as grades 4 versus 5, and 7 versus 8), and classification of student.

First, we placed items on the television monitor with only one problem at a time. We made this decision so that students who needed to follow along with the reading on the television monitor could see each item, even from the back of the room. Also, we based this decision on the findings of Curtis and Kropp (1961) who found that displaying a different number of items on a screen can influence students' performance, in part because of the information that items share. They reported significantly higher scores when one to three items were presented than with the conventional paper and pencil administration (in which many items are presented). To ensure that the problem on the monitor matched the problem in the test booklet and to be certain that students were attending to the problem that was being read, we had the test booklets changed so only one mathematics problem appeared on each page with the facing page blank. Second, we had teachers pace the administration, primarily for logistical reasons, although the findings of Curtis and Kropp (1961) again helped justify this decision. They had found significantly higher performance when items were projected in a paced manner on a large screen (either one or three at a time), relative to taking the test with a traditional booklet and answer sheet. As the test was given in a whole class setting, we had to consider a wide range of student skills. To avoid making the presentation too fast or too slow, we allowed a specific amount of time for students to respond, using the problem difficulty to establish this time. More difficult problems had one minute and easier problems had 30 seconds. Third, we had the answer options "colorized" as they were being read so students could follow along with the various choices. We primarily based this decision on the practical reason that students would be able to see the problems from a distance.

In this study, we employed a specific research design rather than use the program evaluation strategies like those described by Koretz (1997) and Trimble (1998). In the earlier research by Tindal et al. (1998) and Helwig et al. (in press), students were nested in treatments. In contrast, Weston's study utilized a crossed design, in which students received both administration conditions. We also used this last strategy, allowing us to compare each student's performance under the accommodated method of administration to their performance under the standard administration.

PARTICIPANTS

The study was conducted through the Council of Chief State School Officers (CCSSO) as part of the Technical Guidelines for Performance Assessments (TGPA) consortium's research activities. Teachers from 10 states participated (Alabama, Colorado, Connecticut, Delaware, Louisiana, North Carolina, Oregon, Rhode Island, Texas, and

Wyoming). Table 1 displays the total counts for all 10 states. Approximately 2,000 students were tested, evenly distributed by grade with approximately 600 students in fourth and eighth grades, 500 students in seventh grade, and 400 students in fifth grade. We sampled slightly more males than females, particularly in the middle school grades. The majority of students were White (72%), with 16% of the population Black, 7% Hispanic, and 2% Asian/Pacific. A substantial percentage of students received free and/or reduced price lunch (36% total). All but one student communicated with speech. Teachers were told to include students for whom such mathematics testing would be part of the student’s academic program and not to include students with IEPs focused on life skills.

TABLE 1. DEMOGRAPHIC CHARACTERISTICS OF PARTICIPANTS BY GRADES (ELEMENTARY AND MIDDLE) – COUNT (%)*

Variables	Elementary Level		All States Middle Level		Total	
	Count	Percent	Count	Percent	Count	Percent
Grade (total)	988	(100)	1,114	(100)	2,102	(100)
	582 (4 th)	(59)	511 (7 th)	(46)	1,093	(52)
	406 (5 th)	(41)	603 (8 th)	(54)	1,009	(48)
Gender (total)	938	(100)	911	(100)	1849	(100)
Male	479	(51)	490	(54)	969	(52)
Female	459	(49)	421	(46)	880	(48)
Ethnicity (total)	914	(100)	894	(100)	1,808	(100)
White	683	(75)	618	(69)	1,301	(72)
Black	123	(13)	173	(19)	296	(16)
Hispanic	75	(8)	60	(7)	135	(7)
Asian/Pacific	20	(2)	14	(2)	34	(2)
Multi racial	1	(0)	7	(1)	8	(0)
NA	12	(1)	22	(2)	34	(2)
Free lunch (total)	859	(100)	871	(100)	1,730	(100)
Reduced	61	(7)	68	(8)	129	(7)
Free	247	(29)	258	(30)	505	(29)
Neither	551	(64)	545	(63)	1,096	(63)

*Totals may not equal 100 percent due to rounding and totals may not sum across categories due to missing data.

We recruited teachers in targeted grade levels so that in every building we had two general education teachers, each with a classroom of approximately 25-30 students and one special education teacher with all students in that grade level who were receiving services with academically oriented IEPs. In the end, we over-sampled students with disabilities (18% in the elementary grades and 27% in the middle school grades), the vast majority classified with learning disabilities, mental retardation, or speech disabilities. Most students were served in the general education classroom, either with a pull out model for instruction by the special education teacher separately or included within the general education setting. Teachers rated students’ proficiency in English as well above average (96%) or average (3%). See Table 2.

TABLE 2. EDUCATIONAL STATUS OF PARTICIPANTS BY GRADES (ELEMENTARY AND MIDDLE) COUNT (%)*

Categories	Elementary Level		All States Middle Level		Total	
	Count	Percent	Count	Percent	Count	Percent
Classification	927	(100)	1,092	(100)	2,019	(100)
Title 1	144	(16)	121	(11)	265	(13)
SPED	166	(18)	296	(27)	462	(23)
504	18	(2)	24	(2)	42	(2)
ESL/LEP	6	(1)	2	(0)	8	(0)
Regular ed.	593	(64)	649	(59)	1,242	(62)
Types of disability	182	(100)	214	(100)	396	(100)
Mental	4	(2)	15	(7)	19	(5)
Speech	22	(1)	9	(4)	31	(8)
Orthopedic	3	(2)	1	(0)	4	(1)
Traumatic	2	(1)	1	(0)	3	(1)
LD	126	(69)	159	(74)	285	(72)
SED	9	(5)	9	(4)	18	(5)
Hearing	3	(2)	2	(1)	5	(1)
Visual	3	(2)	0	(0)	3	(1)
Autism	1	(1)	0	(0)	1	(0)
Other health	9	(5)	18	(8)	27	(7)
English fluency	934	(100)	885	(100)	1819	(100)
Well above aver.	880	(96)	850	(96)	1730	(96)
Average	27	(3)	30	(3)	57	(3)
Below average	8	(1)	5	(1)	13	(1)

*Totals may not equal 100 percent due to rounding and totals may not sum across categories due to missing data.

About 40% of the students had previously received accommodations in testing. Finally, teachers also rated students' intellectual functioning as average (50% - 53%) or below average (32% - 39%) with few students (8%-18%) rated with retardation.

MEASURES

We administered a multiple-choice mathematics test from one of the participating states, using a fourth grade and a seventh grade level. The test assesses the following seven strands: (1) numeration, (2) geometry, (3) patterns and pre-algebra, (4) measurement, (5) problem-solving, (6) data analysis and statistics, and (7) computation. A sample of 60 items were selected from a larger pool of problems by having state educational representatives review the larger pool and flag all problems that did not appear aligned with their state curriculum frameworks or with actual tests for that grade level in their state. Once the 60 problems were identified, two (alternate) forms were created by matching items, objectives, and proportion correct, using state data from where they originated. See Table 3.

TABLE 3. DESCRIPTIVE STATISTICS FOR PROPORTION CORRECT WITH COEFFICIENT ALPHA FOR FORMS IN ELEMENTARY AND MIDDLE LEVEL TEST.

Proportion Correct	Mean	SD	Min.	Max.	Alpha (n)
Elementary Level					
Form A	.55	.14	.29	.83	.86 (936)
Form B	.55	.19	.20	.90	.84 (913)
Middle Level					
Form A	.40	.16	.11	.78	.79 (921)
Form B	.40	.16	.18	.89	.86 (921)

We counterbalanced the administration of the two measures so that half the students received Form A first and half received Form B first. In this counterbalancing, we also

matched the method of administration so that when students received Form A first, half of the time it was with the video and the other half of the time it was with the standard administration. Conversely, when students received Form B first, half of the time it was with the videotaped administration first and half of the time it was with the standard administration.

We also administered five different criterion measures along with both administrations of the multiple-choice test. These measures also were counterbalanced in administration and used to help understand the reading and mathematics skills of the target populations. All skill measures except the reading task (maze) were read to the students as they read along; the teacher rating measure was administered prior to the multiple-choice testing.

1. A maze task was used to document student reading proficiency. The maze required the student to read a 200-word passage that had only the first (and last) sentence completely intact and 25 words replaced with a blank. One of the passages was written with an elementary and the other with a middle school level readability. Although a typical maze task requires every 5th or 7th word to be deleted, we removed words based on lexical features, deleting words important to the story (nouns, verbs, adjectives, and adverbs). To the right of the passage was a set of 5 choices of words to put in the blank (only one of which was lexically and syntactically correct). The students indicated their choices by filling in the accompanying bubble (A-E) on a response sheet, which was later scanned into a data file. Students were told: *When you come to a numbered blank in a sentence, look to the right and find the numbered group of words. You should circle only one word per blank. Be sure to read all five of the words before you choose. This is not a timed test.* The reliability coefficients (Cronbach's alpha) for these tests were .91 (elementary school version, n = 940) and .92 (middle school version, n = 1009).
2. Sets of 21 (middle school) and 19 (elementary school) mathematics computation problems were administered that contained mixed operations, fractions, and simple story problems and required the students to calculate (not select) the answer. The problems ranged from mathematics facts to problems requiring borrowing and carrying. Two forms were created – one for the elementary and one for the middle school students. Students were told: *Please complete as many of the following problems as you are able to. It is important to show all of your work so that you can receive partial credit if some of your steps are correct. Be sure to watch the signs.* The reliability coefficients (Cronbach's alpha) for these tests were .78 (elementary school version, n = 930) and .84 (middle school version, n = 1028)
3. A 15-item mathematics vocabulary test was used to ascertain students' knowledge of words used in the field of mathematics. All words were selected from elementary school mathematics problems and from middle school mathematics problems for the two levels of the vocabulary test, respectively. These words were displayed on a page with four words listed below them. Students were told: *Pick the one answer that is most closely related to the math vocabulary word in the stem* The reliability coefficients (Cronbach's alpha) for these tests were .56 (elementary school version, n = 862) and .74 (middle school version, n = 1008).
4. Each student also completed 2 constructed-response mathematics problems that had been part of the state-testing program from which the multiple-choice test had been developed. The problems presented information that required a mathematical solution and directed the students to produce an answer and show

their work. The answer was evaluated on a scale of correctness and completeness with a low of 0 and a high of 3. The rater reliability of these problems (percentage of agreement) was 92% and 85% (of 752 raters) perfect agreement for elementary level problems 1 and 2, respectively, and 93% and 96% (of 977 raters for problems 1 and 2 respectively) perfect agreement for the middle level problems.

5. Teachers rated students' mathematics and reading proficiency on a scale of 1-5, corresponding to judgments of "very low," "low," "fair," "high," and "very high," respectively. Teachers were given the following directions: *In the following subject areas, bubble the choice you feel reflects the student's skill. Bubble **only one choice** in each of the subject areas.*

PROCEDURES

Four test forms were created across both grades. For each form of the elementary and middle school test, a booklet was printed in both a standard version (dimensions = 8.5 x 11 inches) and a videotaped version (dimensions = 4.25 x 5.5 inches). The standard version had multiple problems displayed across opposing pages while the video version had only one problem per page on the right side with the opposing (left) page left blank. This system resulted in four different test booklets per grade (Form A—standard and video, and Form B—standard and video). We printed all copies using a color system at each grade so teachers could easily distinguish matching forms (e.g., a green booklet for Form A-video and blue booklet for Form B—standard). Each booklet was marked with an ES (elementary schools) or MS (middle schools) on the cover. All test materials (booklets and directions) were distributed at a training workshop in which teachers received standardized administration directions to read to students. In both conditions, students were allowed the use of a calculator.

For the standard test administration, teachers read the following script:

You are about to take a 30-question math test similar to other math tests you may have taken. You will read each problem and choose the **one** best answer from the four choices given. Your answers should be marked on the separate answer sheet, which is provided. You may use a calculator on any problem you wish. You may also use the paper provided or your test booklet as scratch paper. The test will last for 45 minutes. I will announce when you have 20, 10, and 5 minutes left and 1 minute left.

You are not expected to know how to solve every problem. Some of the questions may involve ideas that you have not talked about yet in your math classes. If you come to a question that you do not know how to solve, work as much of it as you can and then choose which answer you think seems correct. You may go back at any time and change an answer or solve a problem you have skipped.

For the videotaped administration, teachers read the following script:

You are about to take a 30-question math test that may be different from other math tests you have taken. Each question will be read aloud on a video tape by an actor. Each question is also printed in your test booklet exactly as it is being read. You may watch the video monitor as the question is read, or read the question silently to yourself while the actor reads the problem. After the question is read, four possible answers will

be shown on the screen. These choices are also printed in your test booklet.

Choose the **one** best answer from the four choices given. Your answers should be marked on the separate answer sheet that is provided. You may use a calculator on any problem you wish. You may also use the paper provided or your test booklet as scratch paper. Each test problem is printed on a separate page. **Do not** turn the page to the next problem until you are told to do so. You **may not** go back and change answers, so think carefully before you make your choice.

When you are told to turn the page to the next problem you will have 5 seconds to mark your answer (if you have not already done so) and turn the page before the next problem is read.

You are not expected to know how to solve every problem. Some of the questions may involve ideas that you have not talked about yet in your math classes. If you come to a question that you do not know how to solve, work as much of it as you can and then choose which answer you think seems correct.

As each option was read on the video tape, it changed color from white to yellow, and after it was read, returned to white again. After the last item was read, the screen went blank and teachers told the students to “answer the problem now.” Teachers then used an *Individual Item Duration Sheet* that specified how long to pause the video tape. Teachers had the discretion to lengthen the times for each problem (up to twice the length of the suggested times) if they saw that most students were still working on the problem when the suggested time was up. They were told never to shorten the suggested time. During this pause time, teachers could answer individual questions for students or make general comments. When the allotted time was up, students were told to turn the page and the teacher began the video tape. At the bottom of each successive page was printed: **STOP. Do Not Turn The Page.**

Students were assigned an identification number (ID#) prior to participating in the study. This ID# specified the forms, order of the tests, as well as the order for taking the criterion measures. All materials were shipped in time for testing to take place in the period from February through March 1998.

DATA ANALYSIS

All scores for the multiple-choice test were converted to scale scores using the procedures adopted by the state from which they were drawn. For the elementary level, the scale ranged from 121-169 and for the middle level, the range was from 143-193. A total of 243 students were removed from the analysis because of uncertainty in coding the method of administration.

The four criterion measures were converted to z-scores using the mean and standard deviation for the grade level in which they were placed. For example, though both 4th and 5th grade students took the 4th grade level of the mathematics test, the z-score for 4th grades was based on the descriptive statistics for 4th graders only, while the 5th grade statistics were used to compute z-scores for 5th grade students. The same was done with 7th and 8th grade students. All analyses, however, were aggregated as elementary or middle school level.

Three sets of analyses were completed using analysis of variance (ANOVA). First, the effect of administration order was analyzed with a one between (first versus second) and one within (standard versus video). Second, the effect of form was analyzed with a one between (form A versus form B) and one within (standard versus video). Finally, a one between (classification of student), one within (video versus standard) analysis of variance was used to compare students on the treatment variable. The counts across the tables may vary because of missing data. For example, all comparisons for order and form used a nested design and included all students who took any single (list wise) order or form by administration method. These counts, therefore, cannot simply be added together and/or compared directly either to each other or to the demographics (which include only students with no missing data).

First, the effect for the order of the administration was analyzed for each level, elementary and middle. Then the form effect was analyzed. In all of the comparisons, the superscripted number refers to the order of administration (1 = time one and 2 = time two; e.g., SA¹ = Standard administration, form A, time 1 and VB² = Video administration, form B, time 2.).

Order and Form Effects

We conducted four analyses to ascertain whether time of administration (time 1 and time 2) or form (A and B) was confounded with our results. When time effect was analyzed, form was held constant. When form effect was analyzed, time was held constant.

- For the main effect analysis of Time with the standard administration, the two comparisons are: SA¹ versus SA² and SB¹ versus SB². At the elementary level, no order effect was found for the standard administration. See Tables 4 and 5 below.

TABLE 4. DESCRIPTIVE STATISTICS FOR ORDER FROM THE STANDARD ADMINISTRATION FOR ELEMENTARY SCHOOL STUDENTS

Treatment	Count	Mean	Std. Dev.	Std. Err.
Std Admin-Form A ¹	259	147.93	9.13	.57
Std Admin-Form B ¹	219	146.94	10.52	.71
Std Admin-Form A ²	257	147.35	10.17	.63
Std Admin-Form B ²	187	146.55	10.11	.74

TABLE 5. MEAN DIFFERENCES FOR ORDER FROM THE STANDARD ADMINISTRATION FOR ELEMENTARY SCHOOL STUDENTS

Comparison	Mean Diff.	Crit. Diff	P-Value
Std Admin-Form A ¹ vs. Std Admin-Form A ²	.58	2.46	.9323
Std Admin-Form B ¹ vs. Std Admin-Form B ²	.39	2.78	.9851

- For the main effect analysis of Time with the video administration, the two comparisons are: VA¹ versus VA² and VB¹ versus VB². At the elementary level, no order effect was found for the video administration. See Tables 6 and 7 below:

TABLE 6. DESCRIPTIVE STATISTICS FOR ORDER FROM THE VIDEO ADMINISTRATION FOR ELEMENTARY SCHOOL STUDENTS

Treatment	Count	Mean	Std. Dev.	Std. Err.
Video Admin-Form A ¹	187	147.73	10.69	.79
Video Admin-Form B ¹	258	147.35	8.96	.56
Video Admin-Form A ²	217	147.41	9.88	.67
Video Admin-Form B ²	254	148.40	9.06	.57

TABLE 7. MEAN DIFFERENCES FOR ORDER FROM THE VIDEO ADMINISTRATION FOR ELEMENTARY SCHOOL STUDENTS

Comparison	Mean Diff.	Crit. Diff	P-Value
Video Admin-Form A ¹ , Video Admin-Form A ²	-.32	2.69	.9903
Video Admin-Form B ¹ , Video Admin-Form B ²	1.05	2.38	.6748

- For the main effect analysis of Time with the standard administration, the two comparisons are: SA¹ versus SA² and SB¹ versus SB². At the middle school level, an order effect was found with the second administration of the standard form significantly higher than the first administration. See Tables 8 and 9 below:

TABLE 8. DESCRIPTIVE STATISTICS FOR ORDER FROM THE STANDARD ADMINISTRATION FOR MIDDLE SCHOOL STUDENTS

Treatment	Count	Mean	Std. Dev.	Std. Err.
Std Admin-Form A ¹	265	163.01	9.72	.60
Std Admin-Form A ²	252	165.29	9.64	.61
Std Admin-Form B ¹	229	162.84	8.71	.58
Std Admin-Form B ²	151	163.19	9.21	.75

TABLE 9. MEAN DIFFERENCES FOR ORDER FROM THE STANDARD ADMINISTRATION FOR MIDDLE SCHOOL STUDENTS

Comparison	Mean Diff.	Crit. Diff	P-Value
Std Admin-Form A ¹ , Std Admin-Form A ²	-2.28	2.31	.0540
Std Admin-Form B ¹ , Std Admin-Form B ²	-.35	2.75	.9887

- For the main effect analysis of Time with the video administration, the two comparisons are: VA¹ versus VA² and VB¹ versus VB². At the middle school level, no order effect was found for the standard administration. See Tables 10 and 11 below:

TABLE 10. DESCRIPTIVE STATISTICS FOR ORDER FROM THE VIDEO ADMINISTRATION FOR MIDDLE SCHOOL STUDENTS

Treatment	Count	Mean	Std. Dev.	Std. Err.
Video Admin-Form A ¹	152	162.75	8.93	.72
Video Admin-Form A ²	222	163.80	9.02	.61
Video Admin-Form B ¹	254	165.71	10.10	.63
Video Admin-Form B ²	248	163.46	10.02	.64

TABLE 11. MEAN DIFFERENCES FOR ORDER FROM THE VIDEO ADMINISTRATION FOR MIDDLE SCHOOL STUDENTS

Comparison	Mean Diff.	Crit. Diff	P-Value
Video Admin-Form A ¹ , Video Admin-Form A ²	1.05	2.84	.7844
Video Admin-Form B ¹ , Video Admin-Form B ²	-2.25	2.40	.0772

- For Forms A and B with the standard administration: No form effect was found at the elementary level from the standard administration at either time. See Tables 12 and 13 below:

TABLE 12. DESCRIPTIVE STATISTICS FOR FORM FROM THE STANDARD ADMINISTRATION FOR ELEMENTARY SCHOOL STUDENTS

Treatment	Count	Mean	Std. Dev.	Std. Err.
Std Admin-Form A ¹	259	147.93	9.13	.57
Std Admin-Form B ¹	219	146.94	10.52	.71
Std Admin-Form A ²	257	147.35	10.17	.63
Std Admin-Form B ²	187	146.55	10.11	.74

TABLE 13. MEAN DIFFERENCES FOR FORM FROM THE STANDARD ADMINISTRATION FOR ELEMENTARY SCHOOL STUDENTS

Comparison	Mean Diff.	Crit. Diff	P-Value
Std Admin-Form A ¹ , Std Admin-Form B ¹	.99	2.56	.7595
Std Admin-Form A ² , Std Admin-Form B ²	-.80	2.68	.8755

- For Forms A and B with the video administration: No form effect was found at the elementary level for the video administration at either time. See Tables 14 and 15 below:

TABLE 14. DESCRIPTIVE STATISTICS FOR FORM FROM THE VIDEO ADMINISTRATION FOR ELEMENTARY SCHOOL STUDENTS

Treatment	Count	Mean	Std. Dev.	Std. Err.
Video Admin-Form A ¹	183	147.73	10.69	.79
Video Admin-Form B ¹	253	147.35	8.96	.56
Video Admin-Form A ²	217	147.41	9.88	.67
Video Admin-Form B ²	254	148.40	9.06	.57

TABLE 15. MEAN DIFFERENCES FOR FORM FROM THE VIDEO ADMINISTRATION FOR ELEMENTARY SCHOOL STUDENTS

Comparison	Mean Diff.	Crit. Diff	P-Value
Video Admin-Form A ¹ , Video Admin-Form B ¹	.38	2.60	.9821
Video Admin-Form A ² , Video Admin-Form B ²	.99	2.48	.7404

- For Forms A and B with the standard administration: No form effect was found at the middle school level for the standard administration at either of the times. See Tables 16 and 17 below:

TABLE 16. DESCRIPTIVE STATISTICS FOR FORM FROM THE STANDARD ADMINISTRATION FOR MIDDLE SCHOOL STUDENTS

Treatment	Count	Mean	Std. Dev.	Std. Err.
Std Admin-Form A ¹	265	163.01	9.72	.60
Std Admin-Form B ¹	229	162.84	8.71	.58
Std Admin-Form A ²	252	165.29	9.64	.61
Std Admin-Form B ²	151	163.19	9.21	.75

TABLE 17. MEAN DIFFERENCES FOR FORM FROM THE STANDARD ADMINISTRATION FOR MIDDLE SCHOOL STUDENTS

Comparison	Mean Diff.	Crit. Diff	P-Value
Std Admin-Form A ¹ , Std Admin-Form B ¹	.17	2.37	.9979
Std Admin-Form A ² , Std Admin-Form B ²	-2.10	2.70	.1904

- For Forms A and B with the video administration: A form effect was found with Form B significantly higher than Form A at Time 1. See Tables 18 and 19 below:

TABLE 18. DESCRIPTIVE STATISTICS FOR FORM FROM THE VIDEO ADMINISTRATION FOR MIDDLE SCHOOL STUDENTS

Treatment	Count	Mean	Std. Dev.	Std. Err.
Video Admin-Form A ¹	152	162.75	8.93	.72
Video Admin-Form B ¹	254	165.71	10.10	.63
Video Admin-Form A ²	222	163.80	9.02	.61
Video Admin-Form B ²	248	163.46	10.02	.64

TABLE 19. MEAN DIFFERENCES FOR FORM FROM THE VIDEO ADMINISTRATION FOR MIDDLE SCHOOL STUDENTS

Comparison	Mean Diff.	Crit. Diff	P-Value
Video Admin-Form A ¹ , Video Admin-Form B ¹	-2.96	2.76	.0298S
Video Admin-Form A ² , Video Admin-Form B ²	-.34	2.49	.9860

ANALYSIS OF ADMINISTRATION METHOD

The primary analysis of administration method was based on a repeated measures analysis of variance with a subset of two groups of students compared – students with Individualized Educational Programs (IEPs) and general education students; both subsets had been rated low by their teacher on reading proficiency (with a rating of 1 or 2) . A secondary analysis, for both elementary and middle school students, included a population that was expanded to include all students, not just those rated as low in their reading proficiency.

Elementary Level

Two of the three findings for low reading proficiency students are significant. See Tables 20 and 21 below:

TABLE 20. ANOVA TABLE FOR STANDARD AND VIDEO ADMINISTRATION BY ELEMENTARY STUDENT CLASSIFICATION AND LOW READING

	DF	SS	MS	F	P
Classification	1	1602.289	1602.289	16.513	<.0001
Sbj(Grp)	189	18339.585	97.035		
Video-Standard	1	96.639	96.639	4.579	.0336
Video-Standard * Stdnt Class.	1	5.539	5.539	.262	.6090
Video-Standard * Sbj(Grp)	189	3988.419	21.103		

TABLE 21. MEANS TABLE FOR STANDARD AND VIDEO ADMINISTRATION BY ELEMENTARY STUDENT CLASSIFICATION AND LOW READING

Treatment	Count	Mean	Std. Dev.	Std. Err.
General Ed, Standard Admin	112	140.580	7.874	.744
General Ed, Video Admin	112	141.357	7.656	.723
Special Ed, Standard Admin	79	136.177	8.030	.903
Special Ed, Video Admin	79	137.443	7.081	.797

1. Low reading proficiency students (as rated by their teacher) in general education scored significantly higher on the mathematics test than low reading proficiency students in special education with an IEP in reading.
2. For both groups of students, the video administration resulted in significantly higher performance than the standard administration .
3. No significant interactions were found between the two groups of students and the two methods of administration.

The following analysis was conducted with a larger group of students in both general and special education. Instead of confining the population to those rated low in their reading proficiency, all students in general education and those in special education with an IEP in reading are compared.

The findings in this latter comparison (see Tables 22 and 23 below) are similar to those reported when only students with low ratings in reading proficiency were compared. Of course, a significant difference appears in student classification (general education students perform better than special education students with an IEP in reading). Furthermore, the method of administration also reflects higher performance when the test is administered with a videotaped read-aloud over that attained with the student reading it silently.

TABLE 22. ANOVA TABLE FOR STANDARD AND VIDEO ADMINISTRATION BY ELEMENTARY STUDENT CLASSIFICATION

	DF	SS	MS	F	P
Classification	1	30599.146	30599.146	255.340	<.0001
Sbj(Grp)	677	81129.642	119.837		
Video-Standard	1	80.431	80.431	4.962	.0262
Video-Standard * Stdnt Class.	1	48.767	48.767	3.009	.0833
Video-Standard * Sbj(Grp)	677	10972.949	16.208		

TABLE 23. MEANS TABLE FOR STANDARD AND VIDEO ADMINISTRATION BY ELEMENTARY STUDENT CLASSIFICATION

Treatment	Count	Mean	Std. Dev.	Std. Err.
General Ed, Standard Admin	575	151.139	8.317	.347
General Ed, Video Admin	575	151.289	8.311	.347
Special Ed, Standard Admin	104	137.433	8.264	.810
Special Ed, Video Admin	104	138.635	7.455	.731

This effect of the video administration versus the standard administration can also be viewed from an analysis of the entire distribution of students. In Table 24 on the following page, the performance on the standard test administration was subtracted from the video administration. A negative score reveals higher performance with a standard test administration, while a positive score reflects higher performance with the video administration. For elementary students, a slightly greater percentage of general education students performed worse with the video (48%) than the standard version (46%), while for special education students, the opposite was true: More students performed better with the video (53%) than the standard version (40%).

TABLE 24. FREQUENCY DISTRIBUTION OF DIFFERENCE BETWEEN VIDEO-STANDARD ADMINISTRATION BY ELEMENTARY STUDENT CLASSIFICATION

From (≥)	To (<)	Total Count	Total Percent	GEN Count	GEN Percent	SPED Count	SPED Percent
-24	-23	1	0.114	1	0.141	0	0
-23	-22	1	0.114	1	0.141	0	0
-22	-21	0	0	0	0	0	0
-21	-20	1	0.114	0	0	1	0.629
-14	-13	3	0.341	2	0.282	1	0.629
-13	-12	5	0.568	2	0.282	3	1.887
-12	-11	3	0.341	3	0.423	0	0
-11	-10	3	0.341	2	0.282	1	0.629
-10	-9	10	1.135	10	1.41	0	0
-9	-8	18	2.043	16	2.257	2	1.258
-8	-7	13	1.476	11	1.551	2	1.258
-7	-6	18	2.043	16	2.257	2	1.258
-6	-5	42	4.767	36	5.078	6	3.774
-5	-4	38	4.313	31	4.372	6	3.774
-4	-3	74	8.4	64	9.027	10	6.289
-3	-2	60	6.81	53	7.475	6	3.774
-2	-1	60	6.81	47	6.629	11	6.918
-1	0	60	6.81	45	6.347	13	8.176
0	1	54	6.129	44	6.206	10	6.289
1	2	60	6.81	48	6.77	10	6.289
2	3	60	6.81	45	6.347	14	8.805
3	4	40	4.54	32	4.513	7	4.403
4	5	49	5.562	40	5.642	8	5.031
5	6	37	4.2	32	4.513	4	2.516
6	7	27	3.065	20	2.821	7	4.403
7	8	40	4.54	27	3.808	13	8.176
8	9	28	3.178	23	3.244	5	3.145
9	10	19	2.157	13	1.834	6	3.774
10	11	16	1.816	14	1.975	1	0.629
11	12	12	1.362	7	0.987	5	3.145
12	13	8	0.908	6	0.846	2	1.258
13	14	6	0.681	4	0.564	2	1.258
14	15	3	0.341	3	0.423	0	0
15	16	3	0.341	3	0.423	0	0
16	17	3	0.341	3	0.423	0	0
17	18	3	0.341	2	0.282	1	0.629
18	19	0	0	0	0	0	0
19	20	2	0.227	2	0.282	0	0
20	21	0	0	0	0	0	0
21	22	1	0.114	1	0.141	0	0
Total		881	100	709	100	159	100
		Total Count	Total Percent	GEN Count	GEN Percent	SPED Count	SPED Percent
Below 0		410.00	46.54	340.00	47.95	64.00	40.25
At 0		54.00	6.13	44.00	6.21	10.00	6.29
Above 0		417.00	47.33	325.00	45.84	85.00	53.46
Total		881.00	100.00	709.00	100.00	159.00	100.00

Results for totals may not agree with results for individual cells because of missing values for split variables.

Middle School Level

Students in general education who are rated low in their reading proficiency perform better on the mathematics test than students with low ratings in reading proficiency and with an IEP in reading (assuming no interaction of form with method of administration). No effects are present, however, for the method of administration and no interaction exists between classification of student and method of administration. See Tables 25 and 26 below:

TABLE 25. ANOVA TABLE FOR STANDARD AND VIDEO ADMINISTRATION BY MIDDLE STUDENT CLASSIFICATION AND LOW READING

	DF	SS	MS	F	P
CLASSIF.R	1	2195.957	2195.957	26.696	<.0001
Subject(Group)	175	14395.399	82.259		
Video-Standard	1	10.502	10.502	.470	.4937
Video-Standard * Stdnt Class	1	1.146	1.146	.051	.8210
Video-Standard * Sbj(Grp)	175	3907.521	22.329		

TABLE 26. MEANS TABLE FOR STANDARD AND VIDEO ADMINISTRATION BY MIDDLE STUDENT CLASSIFICATION AND LOW READING

Treatment	Count	Mean	Std. Dev.	Std. Err.
General Ed, Standard Admin	99	159.798	7.574	.761
General Ed, Video Admin	99	160.030	8.396	.844
Special Ed, Standard Admin	78	154.667	5.608	.635
Special Ed, Video Admin	78	155.128	6.597	.747

When student groups are not restricted to low reading proficiency (as rated by their teacher), a much larger group is available for analysis. This comparison is simply between middle school students in general education and those in special education with an IEP in reading.

Again, students in general education performed better than special education students with an IEP in reading. No effect was found for the method of administration or in the interaction of student classification and method of administration. See Tables 27 and 28 below:

TABLE 27. ANOVA TABLE FOR STANDARD AND VIDEO ADMINISTRATION BY MIDDLE STUDENT CLASSIFICATION

	DF	SS	MS	F	P
CLASSIF.R	1	21231.093	21231.093	157.068	<.0001
Subject(Group)	622	84076.493	135.171		
Video-Standard	1	24.554	24.554	1.134	.2872
Video-Standard * Stdnt Class	1	1.887	1.887	.087	.7679
Video-Standard * Sbj(Grp)	622	13462.793	21.644		

TABLE 28. MEANS TABLE FOR STANDARD AND VIDEO ADMINISTRATION BY MIDDLE STUDENT CLASSIFICATION

Treatment	Count	Mean	Std. Dev.	Std. Err.
General Ed, Std Admin	513	166.320	9.262	.409
General Ed, Video Admin	513	166.585	9.359	.413
Special Ed, Std Admin	111	155.432	5.877	.558
Special Ed, Video Admin	111	155.901	6.723	.638

As reflected in the repeated measures analysis of variance, the frequency and percentage of students who performed better with the video versus the standard administration was more ambiguous with the middle school than the elementary students. See Table 29 below, which displays the effects of the videotaped administration for both general and special education students. Unlike the elementary population, a greater percentage of general education students performed better with the video (50%) over the standard (44%), while with the special education students the percentages were nearly equal for both administrations (about 47%).

TABLE 29. FREQUENCY DISTRIBUTION OF DIFFERENCE BETWEEN VIDEO-STANDARD ADMINISTRATION BY MIDDLE STUDENT CLASSIFICATION

From (\geq)	To ($<$)	Total Count	Total Percent	GEN Count	GEN Percent	SPED Count	SPED Percent
-27	-26	1	0.118	1	0.162	0	0
-21	-20	1	0.118	1	0.162	0	0
-20	-19	0	0	0	0	0	0
-19	-18	1	0.118	1	0.162	0	0
-18	-17	1	0.118	1	0.162	0	0
-17	-16	2	0.237	1	0.162	1	0.463
-16	-15	3	0.355	3	0.485	0	0
-15	-14	5	0.592	4	0.646	1	0.463
-14	-13	3	0.355	3	0.485	0	0
-13	-12	2	0.237	2	0.323	0	0
-12	-11	9	1.066	7	1.131	2	0.926
-11	-10	11	1.303	7	1.131	4	1.852
-10	-9	20	2.37	13	2.1	7	3.241
-9	-8	15	1.777	13	2.1	2	0.926
-8	-7	16	1.896	9	1.454	7	3.241
-7	-6	27	3.199	17	2.746	9	4.167
-6	-5	35	4.147	28	4.523	6	2.778
-5	-4	41	4.858	32	5.17	8	3.704
-4	-3	29	3.436	19	3.069	10	4.63
-3	-2	57	6.754	36	5.816	20	9.259
-2	-1	38	4.502	27	4.362	10	4.63
-1	0	64	7.583	47	7.593	16	7.407
0	1	49	5.806	36	5.816	13	6.019
1	2	68	8.057	49	7.916	19	8.796
2	3	40	4.739	30	4.847	10	4.63
3	4	53	6.28	42	6.785	10	4.63
4	5	31	3.673	23	3.716	8	3.704
5	6	54	6.398	40	6.462	13	6.019
6	7	16	1.896	13	2.1	3	1.389
7	8	38	4.502	31	5.008	6	2.778
8	9	16	1.896	9	1.454	7	3.241
9	10	24	2.844	20	3.231	4	1.852
10	11	19	2.251	13	2.1	6	2.778
11	12	22	2.607	14	2.262	8	3.704
12	13	3	0.355	3	0.485	0	0
13	14	11	1.303	11	1.777	0	0
14	15	4	0.474	3	0.485	1	0.463
15	16	5	0.592	4	0.646	1	0.463
16	17	3	0.355	3	0.485	0	0
17	18	2	0.237	1	0.162	1	0.463
18	19	2	0.237	1	0.162	1	0.463
19	20	1	0.118	1	0.162	0	0
20	21	0	0	0	0	0	0
21	22	0	0	0	0	0	0
22	23	1	0.118	0	0	1	0.463

23	24	0	0	0	0	0	0
24	25	0	0	0	0	0	0
25	26	1	0.118	0	0	1	0.463
Total		844	100	619	100	216	100

From (≥)	To (<)	Total Count	Total Percent	GEN Count	GEN Percent	SPED Count	SPED Percent
Below 0		381	45.139	272	43.944	103	47.687
At 0		49	5.806	36	5.816	13	6.019
Above 0		414	49.05	311	50.245	100	46.299
Total		844	99.995	619	100.005	216	100.005

Results for totals may not agree with results for individual cells because of missing values for split variables.

RELATIONSHIP WITH CRITERION MEASURES

Four criterion measures were analyzed and are reported in Tables 30 to 37 below. Basic statistics are presented, as well as the intercorrelation among them, for both versions of the multiple-choice test standard and video). The quantitative measures are reported with means and standard deviations while the qualitative measure is reported with frequency distributions. For three measures (mathematics vocabulary, reading maze, and mathematics skill) with elementary school students, z-scores (using grade level means and standard deviations) and raw scores have been reported excluding students classified as Limited English Proficiency with 504 plans.

To help understand the correlational data, descriptive statistics were computed for these criterion measures. For elementary students, the vocabulary and mathematics skills distributions were nearly normal for students in elementary schools, while the distribution for the maze is very negatively skewed and leptokurtic. For the vocabulary test, the average was about 10 (of 15 possible). For the maze test, the mean was 21 (of 25 possible). And finally, for the mathematics skill test, the average was nearly 14 problems correct (out of 21 possible). See Table 30 below:

TABLE 30. DESCRIPTIVE STATISTICS OF THREE CRITERION MEASURES USED AS MARKER VARIABLES IN READING AND MATH SKILLS FOR ELEMENTARY SCHOOL STUDENTS (NO LEP OR 504): Z-SCORES AND RAW SCORES, RESPECTIVELY

Z-scores	Vocabulary	Maze	Math Skill
Mean	-3.26E-5	2.41E-3	3.12E-3
Std. Dev.	1.01	1.00	1.00
Std. Error	.03	.03	.03
Count	920	920	909
Minimum	-3.43	-5.27	-4.25
Maximum	2.17	.81	2.05
# Missing	42	42	53

Raw scores	Vocabulary	Maze	Math Skill
Mean	10.10	21.23	13.75
Std. Dev.	2.40	5.02	3.03
Std. Error	.08	.17	.10
Count	920	920	909
Minimum	1.00	1.00	1.00
Maximum	15.00	25.00	19.00
# Missing	42	42	53

For these three measures with middle school students, the vocabulary measure is far more platykurtic, with some variance, the maze is again leptokurtic and negatively skewed (similar to that found with the elementary students), and the skills test is nearly normally distributed (though slightly negatively skewed). The raw scores for these three measures

also are similar to the results obtained with the elementary students (10 out of 15 for the vocabulary, 20 out of 25 for the maze, and 13 out of 21 for the mathematics skills test). See Table 31 below:

TABLE 31. DESCRIPTIVE STATISTICS OF THREE CRITERION MEASURES USED AS MARKER VARIABLES IN READING AND MATH SKILLS FOR MIDDLE SCHOOL STUDENTS (NO LEP OR 504): Z-SCORES AND RAW SCORES, RESPECTIVELY

Z-scores	Vocabulary	Maze	Math Skill
Mean	3.85E-3	-3.56E-3	-2.28E-3
Std. Dev.	1.00	1.01	1.01
Std. Error	.03	.03	.03
Count	981	983	1002
Minimum	-3.22	-4.41	-2.66
Maximum	1.83	.99	1.89
# Missing	107	105	86

Raw scores	Vocabulary	Maze	Math Skill
Mean	10.25	20.27	13.08
Std. Dev.	3.09	5.04	4.38
Std. Error	.10	.16	.14
Count	981	983	1002
Minimum	0.00	0.00	2.00
Maximum	15.00	25.00	21.00
# Missing	107	105	86

Tables 32 and 33 below, display the correlation matrices for both elementary and middle school students. Most of the correlations among the criterion measures and the multiple-choice test are in the moderate range. The correlation between both forms of the multiple-choice test is relatively quite high.

The correlation among the maze (a reading measure) and the vocabulary (a concept measure) and either the standard or the video version of the mathematics multiple-choice test for elementary students is similar to the correlation between the mathematics skills test and either form of the multiple-choice test. See Table 32 below:

TABLE 32. CORRELATIONS AMONG CRITERION MEASURES AND MULTIPLE-CHOICE TEST (BOTH ADMINISTRATION CONDITIONS & NO LEP OR 504) USING GRADE LEVEL STANDARD SCORES FOR ELEMENTARY SCHOOL STUDENTS

	Vocabulary	Maze	Math Skill	STD.RSCALE	VID.RSCALE
Vocabulary					
Maze	.47				
Math Skill	.36	.40			
STD.RSCALE	.53	.52	.47		
VID.RSCALE	.48	.49	.47	.82	

*881 observations were used in this computation.
81 cases were omitted due to missing values.*

For middle school students, this relationship does not hold up. The highest correlation for these students (beyond the two forms of the multiple-choice tests) is between the mathematics skill measure and either form of the multiple-choice test. The vocabulary measure remains moderately highly correlated with the multiple-choice test (either type of administration) but the maze measure does not maintain the same level of relationship. See Table 33 below:

TABLE 33. CORRELATIONS AMONG CRITERION MEASURES AND MULTIPLE-CHOICE TEST (BOTH ADMINISTRATION CONDITIONS & NO LEP OR 504) USING GRADE LEVEL STANDARD SCORES FOR MIDDLE SCHOOL STUDENTS

	Vocabulary	Maze	Math Skill	STD.RSCALE	VID.RSCALE
Vocabulary					
Maze	.51				
Math Skill	.55	.42			
STD.RSCALE	.54	.37	.56		
VID.RSCALE	.53	.38	.56	.76	

*844 observations were used in this computation.
244 cases were omitted due to missing values.*

For the open-ended mathematics performance task, used with either elementary or middle school students, a difference appears between the two problems with the second one being more difficult (almost half scored zero points in elementary schools and over 80% scored zero points in middle school). See Tables 34 and 35 below:

TABLE 34. FREQUENCY OF RATING VALUES ON OPEN-ENDED MATH PERFORMANCE TASK FOR ELEMENTARY SCHOOL STUDENTS (NO LEP OR 504)

Task 1-Rater 1

Score	Count	Percent
0	143	19.02
1	369	49.07
2	143	19.02
3	97	12.90
Total	752	100.00

Task 2-Rater 1

Score	Count	Percent
0	319	42.42
1	257	34.18
2	117	15.56
3	59	7.85
Total	752	100.00

Task 1-Rater 2

Score	Count	Percent
0	141	18.75
1	372	49.47
2	143	19.02
3	96	12.77
Total	752	100.00

Task 2-Rater 2

Score	Count	Percent
0	322	42.82
1	224	29.79
2	152	20.21
3	54	7.18
Total	752	100.00

TABLE 35. FREQUENCY OF RATING VALUES ON OPEN-ENDED MATH PERFORMANCE TASK FOR MIDDLE SCHOOL STUDENTS (NO LEP OR 504)

Task 1-Rater 1

Score	Count	Percent
0	535	54.76
1	190	19.45
2	146	14.94
3	106	10.85
Total	977	100.00

Task 2-Rater 1

Score	Count	Percent
0	826	84.54
1	125	12.79
2	18	1.84
3	8	.82
Total	977	100.00

Task 1-Rater 2

Score	Count	Percent
0.00	542	55.48
1.00	174	17.81
2.00	151	15.46
3.00	110	11.26
Total	977	100.00

Task 2-Rater 2

Score	Count	Percent
0.00	816	83.52
1.00	130	13.31
2.00	23	2.35
3.00	8	.82
Total	977	100.00

Given the qualitative nature of the measure and the distribution of the scores, the total was calculated by adding the two problems together, which made a scale of 0-12 with two raters summed across both problems; a scale score was then computed and this score was correlated with the standard and video administration of the multiple-choice test. For elementary school students, this correlation was moderately high for the standard (.57) and slightly lower with the video (.50). See Table 36 below:

TABLE 36. CORRELATION FOR ELEMENTARY STUDENTS BETWEEN OPEN-ENDED MATH PERFORMANCE TASK AND STANDARD AND VIDEO ADMINISTRATION OF THE MULTIPLE-CHOICE TEST USING SPEARMAN RANK (NO LEP OR 504 STUDENTS)

Standard Administration of the Multiple-Choice Test and Open-Ended Performance

Sum of Squared Differences	27597191.000
Rho	.571
Z-Value	15.391
P-Value	<.0001
Rho corrected for ties	.565
Tied Z-Value	15.221
Tied P-Value	<.0001
# Ties, OE.SS	13
# Ties, STD.RSCALE	40

234 cases were omitted due to missing values.

Video Administration of the Multiple-Choice Test and Open-Ended Performance

Sum of Squared Differences	32124276.500
Rho	.500
Z-Value	13.493
P-Value	<.0001
Rho corrected for ties	.498
Tied Z-Value	13.429
Tied P-Value	<.0001
# Ties, OESS(Gr)	25
# Ties, VID.RSCALE	39

234 cases were omitted due to missing values.

For middle school students, the same finding appeared: For the standard administration, the two measures intercorrelated .56 and for the video administration, the intercorrelation was .45 as Table 37 illustrates.

TABLE 37. CORRELATION FOR MIDDLE STUDENTS BETWEEN OPEN-ENDED MATH PERFORMANCE TASK AND STANDARD AND VIDEO ADMINISTRATION OF THE MULTIPLE-CHOICE TEST USING SPEARMAN RANK (NO LEP OR 504 STUDENTS).

Standard Administration of the Multiple-Choice Test and Open Ended Performance

Sum of Squared Differences	44446555.500
Rho	.558
Z-Value	16.211
P-Value	<.0001
Rho corrected for ties	.535
Tied Z-Value	15.557
Tied P-Value	<.0001
# Ties, OE.SS	12
# Ties, STD.RSCALE	35

243 cases were omitted due to missing values.

Video Administration of the Multiple-Choice Test and Open-Ended Performance

Sum of Squared Differences	62985636.000
Rho	.447
Z-Value	13.270
P-Value	<.0001
Rho corrected for ties	.440
Tied Z-Value	13.055
Tied P-Value	<.0001
# Ties, OE.SS(Gr)	20
# Ties, VID.RSCALE	39

207 cases were omitted due to missing values.

We hypothesized that students with disabilities should perform better with the use of a videotaped read-aloud than their non-disabled peers. In the elementary level, we found a main effect for both student classification and test administration: Low achieving students outperformed students with IEPs in reading and both groups benefited with a video-tape administration. These same main effects were present when the comparison was all students in general education and all students in special education. In this latter comparison, the interaction we expected was just beyond that expected by chance (.08). For middle school students, although a main effect was found for student classification (general education students outperformed special education students), no main effects for the type of test administration (video versus standard) or interactions were found.

INTERPRETATIONS

Our findings with the elementary students occurred with form or order effects ruled out and is consistent with two other studies done with read-aloud administrations in mathematics multiple-choice tests (Tindal et al., 1998; Weston, 1999).

At the middle school level, though we found no treatment effect, we did find an interaction between form and method of administration. Although we had developed alternate forms of a multiple-choice mathematics test with very comparable difficulty levels (they were similar in mean, standard deviations, and ranges), and we used scale scores in the analyses, we still found differences for middle school students once these tests were administered under the different conditions. This finding may be due to a difference in the population used in establishing the initial levels of difficulty from the population sampled in this study, from the use of a single test across two grade levels, from the use of difficulty estimates that are only based on a standard administration, or from all of these factors combined.

This result may have appeared because of our use of a single test across two grade levels in the study, though we did not disaggregate the scores beyond the middle school level. Furthermore, as noted above, differences in populations between the standardization sample and the study sample may have interacted with grade levels.

Perhaps the best explanation lies in the manner in which problem difficulty is considered, typically on the basis of mathematical operations (their type and fit within a sequence of objectives) and objectively calculated under standard administration conditions. Rarely are text features considered in calibrating items on a scale of difficulty and discrimination; these indices are calculated only with a standard administration and are not based on accommodated administrations that would disentangle (linguistic) features not considered part of the primary construct (mathematics achievement).

One interpretation is that reading may be part of the mathematics construct for either certain types of students or certain types of problems. For example, when Helwig et al. (1999) carefully analyzed the problems on their test, they found levels of success across a broad range of students. For example, in the justification of their study, they cite Mayer (1987), who identified four components of mathematics problem-solving: translation, integration, solution planning, and execution, the first two of which are heavily dependent on reading skill. They also cite the findings of Thompson (1967), who administered two versions of the same mathematics problem-solving task to sixth-grade

students and found that students performed significantly better on the version with lower readability. Finally, they reference the findings of Jerman and Mirman (1974) who reported that the total number of characters, syllables, words, and sentences were related to problem difficulty levels as were word and sentence length. Helwig et al. (1999) reported that: “As the number of verbs present in a passage increased, the difference in success rate in favor of the video accommodation tended to increase” (p.#22). And, although they found no other features of the mathematics problems to be predictive of differential performance with the test administered in a standard fashion or read to students, they did report effects according to student reading proficiencies. When students with low reading skills had difficult reading passages read to them, their performance increased on 5 of 6 questions.

Weston (1999) also reported moderately strong relationships between reading skill of students and performance on multiple-choice mathematics tests, although his results are less easily explained. In a regression plot of student reading performance as measured by the Terra Nova, he found that students with higher reading performance actually benefited from having the test read to them. And a substantial number of students with disabilities did not perform as highly when the test was read to them as when they read it themselves, though, overall, this method of administration resulted in differentially improved performance for students with disabilities on average.

In our results for the middle school students, the form of the test had an impact with one type of administration (the video) and at one time (first administration). Practically speaking, this finding of form interacting with method of administration (and order) means that, for some students, a particular form (A versus B) resulted in a difference of performance on the videotaped administration when it was administered first. In the first administration of the video, Form B was significantly higher than Form A. Clearly, individual problems need to be analyzed for specific linguistic features. However, this analysis needs to be done primarily for students in middle grades. This finding can be related to those reported by Helwig et al. (1999), although he studied 6th grade students while we sampled 7th and 8th grade students. It is possible that this difference in one year is significant in the degree to which mathematics problems become algorithm-based versus linguistic-based, which is not likely to be the case. In the end, with more difficult problems, students will need to know mathematics algorithms, and reading the problems and options is not likely to result in improved performance overall but is effective with specific problems. Reading is necessary but not sufficient for successful solutions. However, we have not analyzed specific problem types as Helwig et al. did in their study.

This interaction was not the one we had anticipated. Rather, we tried to operationalize Phillips’ (1994) contention that validation of an accommodation requires positive effects for those who need it and no effect for those who do not need it. Students with disabilities and reading IEPs were compared to students not being served in any program. We found no such interaction at the middle school level. Rather, the interaction was a method by form effect. This finding may have been because of a novel form of administration with some problems. Quite simply, students were not familiar with a videotaped read-aloud and, although we used a practice item and explained the purpose for this procedure, students with one form did better than students with the other form (but only when it was first administered).

This interaction has two important implications. First, it may reinforce the need for accommodations to be used in the classroom prior to their general use in the testing situation. Second, for students with disabilities, the accommodation must be part of their Individualized Educational Plan.

This area of research appears to have quickly focused on more than simply documenting the effects of an accommodation. With the initial findings of Tindal et al. (1998) reporting on the differential effects of reading mathematics multiple-choice problems, the focus has increasingly shifted from a global consideration of the population to an analysis of the types of problems and now to the interactions of administration methods with problems. No longer is the issue simply one of specifying students as members of disability groups or with IEPs in reading. With the research from Helwig et al. and from this study, our understanding of mathematics achievement testing is beginning to broaden beyond an analysis of problem difficulty in mathematical terms only. Rather, text features may be part of the problem space also. And, the findings from this study indicate that these features may interact with method of administration as well as specific populations, identified by grade level or program participation.

What is needed next is a refined analysis of both form differences and membership categories to address identifying what problems work best with which students as defined by their actual skill levels in reading and mathematics. Only with this type of person-problem fit can we begin to understand the construct validity of both an accountability system and our definition of an accommodation.

REFLECTIONS

Although we conducted an experimental study using appropriate controls for threats to internal validity (limiting any explanations of cause and effect), the study has certain limitations.

Teachers implemented all experimental procedures following a one-day training session, often after a significant delay (some teachers tested within one week and others not until two to three weeks later). We did not monitor teacher implementation and have to assume that they were done properly. We did evaluate, however, the effects of the training at the end of each session and found overwhelmingly positive feedback. Teachers quite uniformly felt capable of running the study in their classrooms, which included scheduling all testing in various classrooms, managing the group administration of the tests, and operating the video tape. A few comments from the training follow:

At this time I'm comfortable with administering the test. I'm looking forward to the results. For myself today's meeting could have been condensed into a half day workshop (study day). I appreciated being asked for input before the tests are finalized.

I feel prepared for the study. The materials were pretty self-explanatory and made things much clearer. I don't think it will be very difficult to administer the tests (and may even help students prepare for other standardized tests). My only concern is rating the students.

I am very interested in accommodating needs of students, so I was very interested in your presentation. You were very organized, and the presentation was well paced.

The sampling plan of the study was neither random nor stratified for teachers or students. Rather, teachers had been nominated for participation based on personal contacts of state department personnel through their own networks with principals and others in the local educational agencies. We cannot ascertain the degree to which the teachers in this study are similar to each other or representative of other teachers in their state. We had only specified that each building should have at least 2-3 teachers in a building (elementary or middle school) within a grade (4 or 5; 7 or 8) and a special education teacher, with two

such buildings selected within either elementary or middle levels. We have not compared teachers on any demographics. We have summarized the key demographics of their students and found considerable differences among the 10 states. Many of these differences, however, have an unknown effect. At the end of each workshop, we collected information about the types of programs and schools in which they worked. For each building, we found great diversity in the configurations for serving students. In the end, our sample of teachers was diverse enough to help us be certain that the findings are not limited to any one subgroup. In a related manner, we have no knowledge of the instructional programs and their alignment with the test. We had asked state department of education representatives to review a large pool of items that could serve as the dependent variable. They were directed to eliminate any items that were NOT aligned with their state standards. Such a review and elimination process, however, does not ensure that teachers taught the skills during the year for solving the problems that remained. For the elementary students, we used a 4th grade test and for the middle school students, a 7th grade test was used. It is likely, therefore, that the items had been taught at some time in the students' school career. Furthermore, being across 10 states, it is likely the variation in content coverage was randomly distributed and not systematically organized. At worst, therefore, this variance becomes part of an error term.

Such limitations notwithstanding, the study controlled for many critical variables by having teachers follow a uniform process for administering and scoring the experimental conditions. Indeed, the study was completed in a more standardized manner than is generally the case with implementation of large-scale assessments. No state can say with certainty that standardization procedures are followed with 100% integrity in all cases: It is assumed rather than documented. The video tape controlled for reading rate (set at 125 words per minute with a teleprompter) and took into account ethnicity and gender by varying the person who read the problems. The readers included an African American, an Asian, and a Caucasian; two readers were female and one was male. The television monitors were new and had large, easily visible screens. Finally, the video player also was new and came with a remote control, allowing teachers to roam among the students and proctor their completion of the test. The booklet was published so that the problem being read was the only one being addressed by the student, ensuring the treatment was being implemented.

Our purpose initially in using a video tape was to not only understand the effect of the read-aloud, but also to provide an accommodation medium that can be used as a part of other accommodations. With videotapes, teachers could access a number of other accommodations, many of which would be quite difficult to implement as needed by students on their caseload. For example, with extended time alone, the most frequently used and investigated accommodation (see Tindal & Fuchs, 1999) would be difficult to provide logistically for a special education teacher with a caseload of 25 students with disabilities. Yet, the use of a videotaped administration would let the teacher accommodate this group by allowing students to "self-administer" with the teacher managing the process. Many other accommodations in setting, time, and scheduling also would be possible with a videotaped administration.

In summary, this accommodation is slightly different in its outcome from that found with other studies, where differential effects were found (Helwig et al., 1999; Tindal et al., 1998; Westin, 1999). Rather, it appears that the use of a videotaped, read-aloud mathematics accommodation has at the very least a small positive effect on some students though it has no differential effect with groups of students. It works better in elementary than middle level grades. At the very least, even in the presence of no differential effects with groups of students, an analysis of the distribution of all scores showed that some students gained. Clearly then, the process is essentially ideographic

(references performance to the individual and previous performance) rather than nomothetic (references performance to a norm group and reflects relative performance) necessitating the need for decisions to be made on an individual basis. This finding is made more certain because of the strong experimental design that was used.

Finally, it should be noted that this accommodation allows even more important accommodations to be implemented, either alone or in combination. Therefore, further research may begin to utilize a videotaped administration along with other, more powerful and individualized accommodations that deal with setting and time.

References

- Curtis, H. A., & Kropp, R. P. (1961). A comparison of scores obtained by administering a test normally and visually. *Journal of Experimental Education*, 29, 249-260.
- Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (1998). *Mathematics test accommodations for students with learning disabilities: Supplementing teacher judgments with the dynamic assessment of test accommodations*. Unpublished manuscript, Nashville: Peabody College of Vanderbilt University.
- Helwig, R., Tedesco, M., Heath, B., Tindal, G., & Almond, P. (1999). Reading as an access to mathematics problem solving on multiple choice tests for sixth grade students. *The Journal of Educational Research*, 93(2), 113-125.
- Jerman, M. E., & Mirman, S. (1974). Linguistic and computational variables in problem solving in elementary mathematics. *Educational Studies in Mathematics*, 5, 317-362.
- Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles: Center for Research on Standards and Student Testing.
- Mayer, R. E. (1987). *Educational psychology: A cognitive approach*. Boston: Little, Brown and Company.
- Phillips, S. E. (1994). High stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, 7(2), 93-120.
- Thompson, E. N. (1967). Readability and accessory remarks: Factors in problem solving in arithmetic. (Doctoral Dissertation, Stanford University, 1967). *Dissertation Abstracts International*, 28, 2464A.
- Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). *Testing accommodations for students with learning disabilities: A review of the literature*. Minneapolis: University of Minnesota National Center on Educational Outcomes.
- Tindal, G. (1997). *Single subject research designs to understand the effects of accommodations in large-scale testing*. Unpublished manuscript. University of Minnesota, National Center on Educational Outcomes.
- Tindal, G. (1998). *Accommodations in large scale tests for students with disabilities: An investigation of reading math tests using video technology*. Unpublished manuscript.
- Tindal, G., & Fuchs, L. (1999). *A summary of research on test changes: An empirical basis for defining accommodations*. Unpublished manuscript. Lexington, KY: Mid-South Regional Resource Center.
- Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study of student response and test administration demands. *Exceptional Children*, 64(4), 439-450.
- Trimble, S. (1998). *Performance trends and use of accommodations on a statewide assessment* (Maryland/Kentucky State Assessment Series Rep. No. 3). Minneapolis,: University of Minnesota, National Center on Educational Outcomes.

Weston, T. (1999, April). *The validity of oral presentation in testing*.: Paper presented at the meeting of the American Educational Research Association; Montreal, CANADA

Ysseldyke, J., Thurlow, M., McGrew, K., & Shriner, J. (1994). *Recommendations for making decisions about the participation of students with disabilities in statewide assessment programs*, (Synthesis Report No. 15). Minneapolis: University of Minnesota, National Center on Educational Outcomes.