EFFECTS OF TESTING ACCOMMODATIONS AND ITEM MODIFICATIONS ON STUDENTS' PERFORMANCE:
AN EXPERIMENTAL INVESTIGATION OF TEST ACCESSIBILITY STRATEGIES

By

Peter A. Beddow

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Special Education

May, 2011

Nashville, Tennessee

Approved:

Professor Stephen N. Elliott

Professor Daniel J. Reschly

Professor Lynn S. Fuchs

Professor Steve Graham

Professor Elisabeth Dykens

TABLE OF CONTENTS

Dedicated to

My brothers: Darin Gordon, Stephen Tedeschi, Scott Hord, Ben Anderson,

and James Smith

(for walking with me through the fire swamp);

Blanca Funes

(for revealing the Lord); and

Mr. Ely

(the best teacher I have ever had).

ACKNOWLEDGMENTS

While there are innumerable people for whom I am thankful, and to whom I will give personal thanks in the months and years to come, I want to acknowledge the contributions of two specific individuals to the work that follows.

The first individual is the One to whom I owe my life: Jesus Christ, who never let go of me even when I let go of Him; and His Body the Church, in whose arms I rested when I became weary. It is by the grace of God that I am sustained, and it is from Him that all of my gifts have proceeded.

Second, this project and my related academic degree would not have come to fruition without the support, encouragement, and faith of Steve Elliott. Steve has guided my doctoral work with care and unwavering integrity, two qualities without which I likely would not have succeeded, and certainly would not have thrived.

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Concrete Statement of the Problem

The current investigation consisted of an examination of the effects of two methods for increasing test accessibility on the test performance of students with a broad range of abilities and needs. Results are used to refine accessibility theory, to advance the development of accessible tests, and to improve measurement of achievement for all students. *Test accessibility* is defined as the extent to which a test event permits a test-taker to demonstrate his or her knowledge of the target construct (Beddow, Elliott, & Kettler, 2009a). Thus, an accessible test or test item presents no construct-irrelevant barriers that prevent the test-taker from showing the extent to which he or she possesses the knowledge, skills, or abilities measured by the test. To the extent a test demands physical, material, or cognitive resources in excess of the construct it is designed to measure, inferences made from the scores on the test are more likely to reflect in some part the accessibility of the test. The implications of such test accessibility concerns are salient particularly for test-takers for whom extraneous test or item demands preclude them from demonstrating what they know. In essence, extraneous demand reduces a test's accuracy and precision as a measuring tool for students for whom extraneous demand poses a hindrance, while test accessibility is not reflected in the inferences made from test scores for students for whom the extraneous demand does not reduce the accessibility of the test. In recent years, test access concerns have been addressed in a number of ways, including using testing accommodations (e.g., Sireci, Scarpati, & Li, 2005) and more recently through the application of universal design

principles in the test development process (Johnstone, Thurlow, Moore, & Altman, 2006). I will hereotofore discuss these two general access strategies using the terms *testing accommodations* and *test modifications*. Testing accommodations refer to strategies aimed at increasing access during testing and involve changes in test administration procedures, whereas test modifications occur prior to the test event and involve changes to the test itself (typically during the development of test items and test forms.)

The primary goals of the present study were threefold. The first goal was to examine the relative and additive effects of testing accommodations and test item modifications on the test performance of a diverse sample of seventh-grade students. The second was to examine the relations of test accessibility with common psychometric indices. The third was to examine students' perspectives about access strategies and related issues. The study was guided by universal design principles, cognitive load theory, and professional testing standards and is part of a continuum of programmatic research on accessibility and the validity of inferences from test scores for all individuals, particularly those for whom tests typically have posed major challenges.

Guiding Questions

Four fundamental questions shaped the design and focus of this study. These questions were the following:

1. What are the effects of testing accommodations, item modifications, and a combination of the two on students' test performance?

2. Are there greater score boosts across the range of test accessibility methods for students with IEPs (individualized education programs) compared to students with no IEPs?
3. What are the relations among accessibility and other test item characteristics?
4. How do students perceive access to learning and testing across the educational arena and what are their reactions to strategies to increase test accessibility?

The study was influenced by, but was independent of, a federally-funded project providing states technical assistance to develop and validate modified alternate assessments for students with disabilities.[1]

Theoretical Statement of the Problem

The central concern of the current study involved the effects of strategies to increase access and reduce construct-irrelevant variance, thus improving the validity of inferences made from test scores. Validity, as defined by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) is "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests….[and] is, therefore, the most fundamental consideration in developing and evaluating tests"(p. 9). Tests themselves are not subject to the validity question; rather, validity is attributed to the *inferences* made from the scores yielded by tests to the extent those inferences are supported by evidence and theory. When scores from a particular test are used to generate more than one type of inference, each type must be evaluated. Thus, if tests yield scores from which different inferences are made for different populations, each inference must be validated.

In the arena of achievement testing, manifold factors influence the validity of inferences made from test scores. Among these is the universality of access permitted by a test to all of its components and features, including the construct the test is designed to measure. This test characteristic is called *accessibility* (Beddow, Kettler, & Elliott, 2008). Conceptually, accessibility is proportional to the validity of inferences made from achievement test scores. For a test user to assert an inference about test-taker achievement, the test score must reflect a measure of the intended construct that is free from error. Valid inferences made from achievement test scores must be grounded on the assumption no differences existed across the tested population regarding the ability of test-takers to perform the skills necessary to permit measurement of the intended construct (i.e., access skills). To the degree incomplete access precluded a test-taker from fully engaging the test to the degree assumed by the inferences made from his or her test score, the inferences are invalid for that test-taker.

Related to the need for access across the range of test-taker needs and abilities is the concept of *universal design*, typically defined as the development of products and services that are usable for the entirety of the population for whom they are intended. The Center for Universal Design (CUD; 1997) lists seven primary aspects of universal design: (a) equitable use, (b) flexibility in use, (c) simple and intuitive use, (d) perceptible information, (e) tolerance for error, (f) low physical effort, and (g) size and space for approach and use. Initially, these principles primarily were applied to architectural features to ensure the full range of the population have complete access to buildings and the services contained therein "without the need for adaptations or specialized design" (e.g., ramps and hallways to accommodate individuals who use wheelchairs, Braille lettering on signs, adjusted heights for drinking fountains; CUD, 1997), but they have application to other areas. Specifically, the last decade has

seen a shift in focus toward applying universal design principles across the educational arena, called universal design for learning (UDL; e.g., see Rose & Meyer, 2006).

Universal design is now integrated in federal legislation and has been applied variously to the areas of education and, more recently, to the evaluation of student learning. To wit, the National Center on Educational Outcomes (NCEO) released a number of technical reports include general suggestions for ensuring tests and items adhere to universal design principles (e.g., Thompson, Johnstone, & Thurlow, 2002; Johnstone et al., 2006). Additionally, the National Accessible Reading Assessment Project (NARAP) released a document representing the efforts of a number of testing experts and test companies to guide developers of reading tests (Thurlow et al., 2009). Taken together, these documents represent a much larger discussion about applying universal design principles to the design of assessments to permit increased access and yield better measurement for more students.

Insofar as accessibility is a characteristic of the test event and not solely of the test-taker or test user, it is therefore dependent both on individual differences among test-takers and on factors controlled by test users (the notion of accessibility as it regards the test event will be discussed in detail later.) As it regards the former, test-taker access may differ across individuals depending on differences and needs. Winter, Kopriva, Chen, and Emick (2006) defined *access* as "…the interaction between construct irrelevant item features and person characteristics that either permits or inhibits student response to the target measurement content of the item"(p. 276). In one instance, a test may be maximally accessible to the majority of test-takers, but be inaccessible to the balance of test-takers who are blind. Another test may be maximally accessible to most test-takers, but be largely inaccessible for individuals who are unable to hold a writing instrument or use a computer keyboard. In both of these cases, test developers and users

(e.g., test administrators) may increase the accessibility of the test by altering the administration or response conditions of a test to accommodate the needs of test-takers for whom the standard test conditions do not permit complete access. In many cases, if test users select and use accommodations appropriately and effectively, subsequent inferences made from test scores do not reflect error that is the result of the interaction between the test-taker's individual needs and the test itself.

The use of non-standard changes in testing conditions, such as the prescribed changes involved in testing accommodations for students identified with disabilities, while they are needed to moderate the negative effects of individual test-taker needs on test results, increase the potential for reducing the validity of inferences from the test results. According to the *Testing Standards,* "Each step toward greater flexibility almost inevitably enlarges the scope and magnitude of measurement error. However, it is possible that some of the resultant sacrifices in reliability may reduce construct irrelevance or construct underrepresentation in an assessment program" (AERA, APA, and NCME, 1999, p. 26). Thus, while the expectation that an assessment could be designed such that no individual accommodations are necessary is implausible, an ideal assessment is one that can be administered under the same conditions across the population to ensure resulting inferences are equally valid across the range of that population.

Notwithstanding the compelling rationale for test developers and test users to attend closely to test accessibility in an effort to ensure inferences based on achievement test scores are valid, the construct of test accessibility represents a convergence of several attributes of tests and test items and until recently it had not previously been measured or quantified. Until recently, little effort had been made to evaluate test accessibility with the goal of developing achievement

tests that yield scores from which inferences can be made that do not reflect error resulting from incomplete test-taker access. The present project contributed to the research on access across the scope of the educational arena by introducing a means by which to improve test accessibility. Additionally, the project permited the specific examination of the test performances of students from which the validity of subsequent inferences may be suspect due to accessibility concerns (e.g., for students with special needs or students identified with disabilities).

It is helpful to situate test accessibility within a unified model of educational access (see Figure 1.) In the context of education, there are at least two primary points at which student (or test-taker) access is a concern for learning and testing: (a) at the level of the school or classroom, and (b) during the test event. At the level of the school, concerns about incomplete access are operationalized in the student's opportunity to learn the general curriculum and are addressed through changes in school- or classroom-level variables such as curriculum and instruction. Access concerns at the level of the test are operationalized in test accessibility and are addressed either by modifying the test conditions (testing accommodations) or the test itself (test and item modifications). When achievement tests that are intended to measure content mastery are given to students who have not had adequate access to the curriculum, subsequent test score inferences are invalid. Similarly, tests that are not accessible yield scores from which subsequent inferences reflect both the intended (target) construct of the test and ancillary requisite constructs (ARCs). In both cases, incomplete access results in invalid test score inferences and, likely, decisions based on misinformation. Thus, testing accommodations and test modifications may be used individually, or in tandem, to increase access to the target construct of a test.

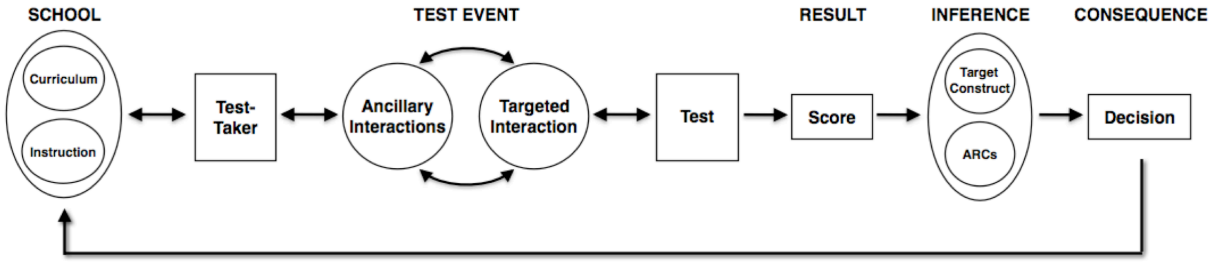


Figure 1. *Unified model of educational access.*

Rationale for the Invention of the Problem

Three aspects contributed to the rationale for the invention of the present research problem of how to develop accessible assessments that yield better measurement of achievement for all students. The first was a *legislative* rationale based on federal requirements for universal participation in assessment, including the impetus for states to develop tests for students for whom current assessment results do not permit valid inferences about their achievement. The second was a *validity* rationale based on the need to ensure test scores reflect accurate and precise measurements of student achievement that can be used to document achievement status and progress. Finally, there was a rationale based on *existing research* on methods for enhancing test accessibility for the target test-taker population.

Rationale Based on Federal Legislation

Universal assessment participation. Under current federal law, states are required to report adequate yearly progress based on results of state and district assessment systems or face serious fiscal and managerial consequences. States also are required to report rates of student participation in these assessment systems across the population, including students identified with disabilities. The Individuals with Disabilities Education Act (P.L. 105-17; 1997, 2004)

contained requirements that states: (a) ensure all students, including students identified with disabilities, participate in all state and district assessments (with appropriate accommodations as needed); and (b) develop alternative assessments to permit the participation of students who are unable to participate in general state and district assessments. Under IDEA, states are required to demonstrate that these alternate assessments are "aligned with the State's challenging academic content standards and challenging student academic achievement standards" (§612(a)(16)(C)(ii)(I).

Universal design. Additionally, IDEA (2004) required all state and district-wide assessments to adhere to principles of universal design. Current federal legislation requires the application of universal design principles to the development of all state and district-wide achievement tests. As defined in the Assistive Technology Act (P.L. 105-394, 1998), universal design is "a concept or philosophy for designing and delivering products and services that are usable by people with the widest possible range of functional capabilities, which include products and services that are directly usable (without requiring assistive technologies) and products and services that are made usable with assistive technologies" (§3(17)). While the term accessibility is not used in this definition, universal design principles as applied to assessment technology clearly are intended to address issues of access while responding to the concern raised in the Testing Standards that the use of individualized accommodations may increase measurement error. To wit, this legislation provides the rationale for the use of universal design principles as follows:

> The use of universal design principles reduces the need for many specific kinds of assistive technology devices and assistive technology services *by building in accommodations* for individuals with disabilities before rather than after production. The

> use of universal design principles also increases the likelihood that products (including services) will be compatible with existing assistive technologies. These principles are increasingly important *to enhance access* to information technology, telecommunications, transportation, physical structures, and consumer products ((PL105-394(§3(10)); emphasis added).

Alternate assessments. Commensurate with IDEA (2004) requirements, the initial form of the No Child Left Behind Act (NCLB; P.L. 107-110; 2001) permitted states to report proficiency for calculation of state adequate yearly progress (AYP) using alternate assessments of alternate achievement standards (AA-AASs) for a small percentage of students identified with significant cognitive disabilities for whom the general assessment is not appropriate. Notwithstanding, substantial evidence indicated among the balance of students (i.e., those not participating in an AA-AAS), there were a large number for whom current assessments did not provide results from which valid inferences could be made about their achievement, even with properly implemented and appropriate accommodations. In 2007, subsequent regulations under NCLB were released to address this problem. These regulations permitted states to develop an alternate assessment based on modified academic achievement standards (AA-MAS), the results of which can be used to report up to 2% of the population in a state proficient in a given content area, with the possibility of reporting up to an additional 1% if the percentage of the state's population taking the AA-AAS is short of the 1% cap, up to a maximum of 3% (e.g., if proficiency is reported from only .5% of the student population based on the AA-AAS, proficiency may be reported for up to 2.5% of students based on the AA-MAS.) NCLB requires that all alternate assessments must undergo a rigorous technical review process.

In essence, any state that opts to include an AA-MAS in its accountability system is afforded the opportunity to develop an assessment that reduces the need for individualized accommodations for students identified with disabilities while simultaneously reducing construct irrelevance and construct underrepresentation. Given the emphases in both IDEA and NCLB on universal design for assessment and ensuring the validity of inferences across the population, it is essential that states apply evidence-based principles of test design for the purpose of maximizing the accessibility of tests for all students, including, but not limited to, AA-MASs for students identified with disabilities.

Rationale Based on Validity of Measurement and Instruction

Students across the United States (and indeed, across the globe) participate in achievement tests. To the extent test developers fail to address the issue of the accessibility of tests and test items, the inferences made from scores on their tests are suspect. This concern is of essence for test-takers as important decisions are made based on these inferences regarding curriculum, instruction, placement, and advancement. For students who tend to perform well on assessments and for whom assessments do not pose limits on opportunities for academic success, the accessibility of most tests and test items likely is sufficient. For students for whom assessments pose significant challenges and for whom scores typically result in inferences reflecting negative attributions of these students' achievement or their teachers' competence, it is essential scores do not reflect error that is the result of an inability to access content requisite for demonstrating performance on the test. To the extent scores reflect unaddressed barriers to access, test developers and test users may be culpable for unintended negative consequences for test-takers and for those who use the resulting information.

Underlying the question about the relation between accessibility and unintended consequences for test-takers in the current climate of educational accountability, however, is a deeper concern about the measurement of student achievement. The two primary goals for the use of any measuring tool are the accuracy and precision of the resulting measurements. Based on the degree of these attributes, the user can be confident about his or her use of the measurements as supportive evidence for subsequent decisions. This is equally true for achievement testing as it is with carpentry, engineering, psychology, and so forth. In the case of achievement testing, incomplete test-taker access likely will result in negatively-biased (i.e., inaccurate) scores because the test-taker does not have the full opportunity to demonstrate performance on the target construct. Moreover, intrinsic sources of error (i.e., to the test or the items) apart from error due to accessibility are likely to be magnified because of the barriers posed by access issues that must be negotiated for the test-taker to engage the target construct, resulting in decreased precision. Indeed, an inaccessible test or test item measures unintended knowledge, skills, or abilities apart from the targeted construct. Thus, the measurement validity of the item is conditional on mastery of these other constructs. In the development of accessible test items, each of these skills should be addressed as an additional requisite construct (ARC) that is necessary for responding for some portion of the test-taker population.

Testing accommodations. Even when tests are designed or modified such that accessibility concerns are reduced, there likely are some students for whom the test event continues to yield scores that reflect the measurement of ARCs. In the effort to ensure fairness and inferential validity for students with special needs, federal law permits the alteration of test administration or response features of a test for some students. These are referred to as testing accommodations, and are a common strategy for increasing test accessibility for some students

on an individual basis. Thus, they lie somewhere on the continuum of accessibility enhancement strategies (along with item modifications, which are discussed in the following section on accessibility theory.) Due to the individual nature of their use, the Testing Standards has implied that non-standardized changes to tests should be avoided when possible; in fact, Sireci and colleagues noted that the notion of an "accommodated standardized test" is itself oxymoronic (p. 457). Notwithstanding, testing accommodations historically have been widely used with the aim of reducing construct-irrelevant variance due to the access skill deficits of individual test-takers.

As described by Hollenbeck, Rozek-Tedesco, and Finsel (2000) and Sireci et al. (2005), testing accommodations typically involve changes in the *presentation* of a test (e.g., oral delivery, paraphrasing, Braille, sign language, encouragement, permitting the use of manipulatives), the *timing* of a test (e.g., extended time, delivering the test across multiple days), the *mode of response* (e.g., permitting test-takers to respond in the test booklet instead of on the answer sheet, transcription) or the *environment* (e.g., separate room, elimination of distractions).

Hollenbeck et al. (2000) identified four attributes of appropriate testing accommodations: (a) unchanged constructs, (b) individual need, (c) differential effects, and (d) sameness of inference. In essence, the authors posited that appropriate accommodations, while applied individually based on specific test-taker needs, should not interfere with the test's measurement of the target construct and should permit the same validity of inferences from the results of the test as those from unaccommodated students. Further, the authors argued that the application of accommodations should differentially affect test results for test-takers for whom the accommodations are intended compared to those for whom testing accommodations are not needed. Specifically, for students who need them, resulting scores with accommodations should be higher than without, but for students who do not need accommodations, scores should be the

same in both conditions. This has been referred to as the *interaction hypothesis* (Sireci et al.). For Hollenbeck and colleagues as well as others, this interaction (which is sometimes referred to as a *differential boost*; Fuchs & Fuchs, 2001) is an essential aspect of an appropriate accommodations. At issue is equity for all students. If evidence indicates that testing accommodations may not only boost the test scores of test-takers who are eligible for accommodations but also the scores of ineligible test-takers, it can be argued the inferences from scores of unaccommodated students are negatively biased. This is of particular concern with regard to the lowest-performing students who have not been identified for special education and/or currently do not receive accommodations.

Indeed, in their National Research Council-commissioned review of research on testing accommodations, Sireci and colleagues (2005) indicated this primary intended result of testing accommodations may not be manifest in practice. The authors reviewed 28 experimental, quasi-experimental, and non-experimental empirical studies on the effects of testing accommodations over nearly two decades . They found the most common accommodations were reading support (39%) and extra time (24%). Aggregate results of studies on reading support (usually in the form of verbatim presentation of directions and test items) were mixed. For five of the six studies of the effect of the accommodation on scores from mathematics tests, the interaction hypothesis was upheld. For two studies on reading and two studies across multiple content areas, the interaction hypothesis was not upheld. The authors concluded reading support, while likely increasing the validity of inferences for mathematics tests, may not have the desired effect when used with tests of other content domains. Results of five out of eight studies on extended time indicated students identified with disabilities exhibit higher score gains than students not identified with disabilities when given extra time. The results of one study rejected the

interaction hypotheses, and the results of two other studies did not indicate extra time resulted in gains for either group. Based on these findings, Sireci and colleagues concluded that while the interaction hypothesis was not strictly upheld, "evidence…is tilted in that direction"(p.469). Sireci and colleagues also reviewed several studies on the effects of multiple accommodations (i.e., accommodation packages). The findings of the four studies that used experimental designs supported the interaction hypothesis.

Reported effect sizes of these studies appear small, but there is evidence they may be practically significant. In a survey of accommodations literature, Kettler and Elliott (in press) reported in some studies, effect sizes from accommodations for students with IEPs were twice those for students without IEPs. In one study, effect sizes ranged from .13 for students without IEPs to .42 students with IEPs. While conventional interpretations of effect sizes (e.g., Cohen, 1988) would suggest these effects are statistically unimportant, a meta-analysis conducted by Bloom, Hill, Black, and Lipsey (2008) may provide evidence to the contrary. Bloom et al. found that mean effect sizes of achievement gains across six standardized achievement tests from the Spring semester of one school year to the next range from .06 to 1.52 in reading and .01 to 1.14 in mathematics, with larger effect sizes consistently observed for lower grades and steadily decreasing until grade 12. Further, the data suggest a steep drop in effect sizes from grade K until grade 5, after which no effect sizes above .41 are observed for either reading or mathematics through grade 12. This indicates effect sizes of .40 or higher for students with disabilities may reflect a practically significant intervention from testing accommodations. Indeed, the differential boost reported by Kettler and Elliott provides evidence of an interaction that may heretofore have been underestimated. As applied to the accommodations literature, these results

suggest for some students, appropriate accommodations may indeed reduce barriers and yield more accurate measures of achievement.

Results of a recent study by Feldman, Kim, and Elliott (in press) indicated that in addition to their effect on student scores, testing accommodations increased student test self-efficacy and motivation. Indeed, while the authors found a significant boost, and did not find a significant interaction of test scores between special education and general education students (i.e., test scores increased equally across groups), data showed a significantly larger increase in self-efficacy and motivation for students identified with disabilities compared to their non-identified peers.

Nevertheless, there are a number of challenges associated with implementing testing accommodations. First, many students, particularly adolescents, are averse to them for several reasons, including the fact that the accommodations often draw attention to student challenges (e.g., when a test is read aloud to the student by a teacher or other adult). In practice, students may not avail themselves of the potential support of accommodations even when it is available to them.

Additionally, there are logistical challenges associated with the appropriate implementation of testing accommodations; including time, personnel, and cost, which often result in poor integrity. Another challenge is the difficulty in identifying which students should receive specific accommodations, and which combination of accommodations may be appropriate. Further, little is known about the extent accommodations interact with each other differentially across students or packages, notwithstanding the breadth of the research base on their use. Finally, each time accommodations are used, general and comparative validity are

threatened. Not only is a variable introduced into the test event with each accommodation, but it is introduced for some students and not all.

Opportunity to learn. Even an optimally-designed test with effective accommodations for those who need them will not permit students to perform well if they have not received instruction in the tested content. Regarding assessment of student achievement, Haertel and Calfey (1983) wrote, "And here is the crux of the matter, for the only guide we can use to establish the validity of a test is the validity of the instructional program itself" (p.126-127). While the concept of opportunity-to-learn (OTL) has been explored in educational research and policy for more than 30 years, few researchers have investigated the topic in the context of statewide assessment programs. Notwithstanding, considerations of OTL raise great concerns about the fairness and equity of these and other tests and about the various high-stakes decisions that are connected to their outcomes. Regarding OTL, Porter (1993) asked the question, "Would it be fair to hold students accountable for knowledge they have not been given a fair opportunity to learn?" (p.21)

Legislative mandates have required schools to provide all students, to the extent possible, access to the general education curriculum (IDEA, 1990; 1997). The persistent focus of legislative policy on inclusion for students identified with disabilities in curriculum clearly values the necessity of ensuring equal educational opportunities for all individuals. Further, the commensurate requirement that all students participate in assessments of grade-level content implies schools have responded accordingly by addressing the concern. Indeed, the very requirement that all students be included in assessment-for-accountability programs for reporting AYP assumes the legislation has achieved its objective. Without being certain that schools are providing instruction that is aligned with the content standards, however, assessing students'

proficiency on the standards is likely to yield results from which few inferences can be made with confidence. Indeed, the confidence with which accountability decisions can be made based on inferences about student mastery of the content standards, which in turn are based on results of a test, is inversely proportional to how much is known about the quality and alignment of classroom instruction to the tested content.

Numerous efforts have been made to develop accurate measures of OTL and alignment, but no researchers have reported investigations of the relations among OTL, test accessibility, and test performance. Several investigators have examined the relations among teacher- and student-reported OTL and student achievement, including projects within the International Association for the Evaluation of Educational Achievement (IEA), several curricular alignment studies conducted in the 1970s and 80s (e.g., Borg 1979), and, more recently, studies conducted by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST; e.g., Herman, Klein, & Abedi, 2000). Findings from these studies, coupled with new approaches to solving the problem of OTL and assessment, have resulted in advances in curricular alignment technologies that facilitate the measurement of OTL across the scope of content areas and grade-levels. Clearly, the field has reached a point where assessment researchers must determine to employ these new techniques to conduct rigorous examinations of the relation between OTL and testing, and policymakers and test developers alike must attend carefully to the resulting data. The validity of large-scale assessment depends on it.

Rationale for the Solution to the Problem: Accessibility Theory

The purpose of any test is to gather information about the extent to which a test-taker possesses, or lacks, the target construct intended to be measured. In the case of an achievement

test, the construct typically involves a domain of knowledge, a skill, or an ability. Accessibility theory provides a framework for improving tests for all individuals by offering a perspective on the measurement of this target construct in terms of three sets of variables: the test-taker, the test, and the test event (see Figure 2). A *test event* is the interaction between the test-taker and the test. The test should generate a test event that facilitates the test-taker's interaction with the construct such that the event yields accurate, consistent, clear, and useful information about the amount of the construct the test-taker possesses.

To the extent a test-taker is able to interact with the test in such a way that the event yields valid information about his or her amount of the target construct, the test event is optimal. If the test event is suboptimal, however, the test itself may not be the sole cause of the measurement deficit, and the use of the term *accessibility* to encapsulate this deficit is only apropos insofar as it can be determined that the source of the test-taker's inability to interact with the test is intrinsic to demands of the test that are extraneous to the target construct. If so, then the issue is one of accessibility.

Each individual approaches a test event with specific abilities and limitations. The purpose of the test is to measure one of these, or a set of these, *to the exclusion of the rest*. To the degree individual characteristics other than the individual's amount of the measured construct interact with aspects of the test, resulting test scores may yield invalid inferences about the person's level of the targeted construct.

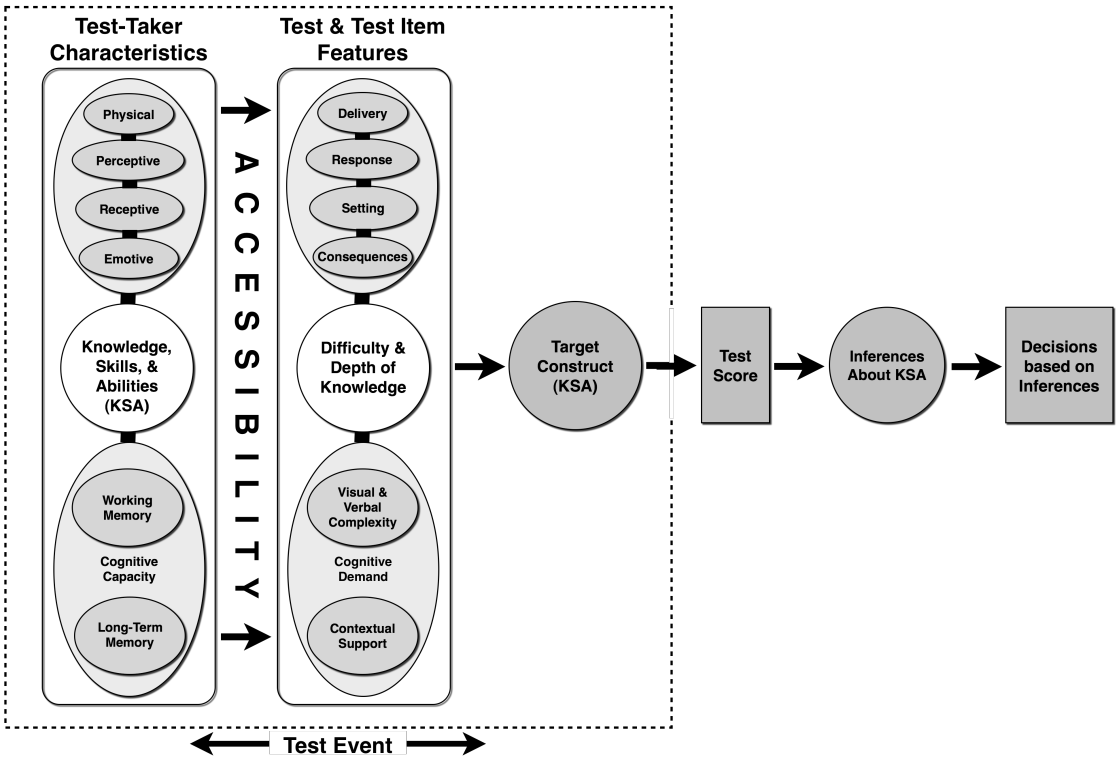


Figure 2. *Test accessibility theory.*

An optimal test event, therefore, is one that yields scores that reflect only the interaction between the individual's level of the target construct and the test itself. To ensure this interaction is unadulterated by other individual characteristics, the test should be designed in such a way that any and all *other* interactions are accounted-for, and their influences on the test score – and thus, subsequent inferences – are reduced to nil. If the test event involves interactions between ancillary individual characteristics and test demands and scores are not properly adjusted to account for these interactions, subsequent inferences about the amount of the target construct the test-taker possesses likely are invalid. While the information provided by the test event may reflect in some part the amount of the intended construct, it also contains information about the test-taker's ancillary characteristics, which also interacted with the test. In essence, not only does

the test measure the target construct, but also it measures what are referred to as ancillary requisite constructs (ARCs; see Figure 1).

Few, if any, tests represent pure measures of their intended constructs. By virtue of the fact that a test is required to gather information about the target construct, an interaction between the test-taker and the test, and the consequent occurrence of a test event, is necessary. It is rare for a test to provide a direct measure of the target construct. Such a test may be termed an *immediate test*: one that requires no mediating event to yield data about the target construct. Theoretically, an immediate test requires no interaction between the test subject and the test itself – essentially unifying the test event with the test, providing a direct measure of the target construct. While certain medical tests may appear to be immediate tests, upon consideration it becomes apparent that even these may be mediated by a test event. For instance, a radiographic ("X-ray") scan may appear to be an immediate test of the composition of tissue and bone. Notwithstanding, similarities between tissues of the body may present complications on an X-ray photograph, producing results from which inaccurate inferences sometimes are made. This type of complication represents an interaction between the subject and the test itself and demonstrates that even what appears to be an immediate test may indeed be mediated by a test event.

Issues of accessibility, by contrast, involve a particular type of interaction: test-taker participation. Questions of accessibility arise only for tests that require the test-taker to interact intentionally with the test (i.e., to respond to a stimulus). Accessibility, therefore, does not belong to the test, nor does it belong to the test-taker. Accessibility is an attribute of the test event.

The implications of locating accessibility in the test event are threefold. First, this accounts for the individual characteristics the test-taker brings into the test, some of which may

interact with the test. Second, locating accessibility in the test event accounts for the features of tests which inadvertently may interact with these individual characteristics. Third, it accounts for the reason testing accommodations and test modifications may be used individually or in tandem to reduce the influence of ancillary interactions (discussed in detail in the next section), thus increasing test-taker access. Figure 2 illustrates a number of interactions between a test-taker's characteristics and features of the test.

To advance understanding of Figure 2, let us examine a hypothetical achievement test called Test A and apply cognitive load theory. Test A was designed to measure a specific hypothetical achievement-related construct, Construct C. Let us assume that Test A, like many achievement tests, is a paper-and-pencil test. Test A generates a test event when the test-taker interacts visually with the test (i.e., via the test items). Similarly, if Test A were presented auditorily (i.e., via sound alone), the test-taker would be required to interact with the test by hearing. These types of interactions are referred to as *perceptive interactions*. Moreover, Test A not only requires the test-taker to perceive a stimulus, but also it demands the test-taker actively receive and interpret the stimulus to deliberately process it in his or her mind before responding. These types of interactions are called *receptive interactions* and include reading and listening.

Once the stimulus is perceived and received, the test-taker must use his or her cognitive resources to process the stimulus. According to cognitive load theory (CLT), cognitive processing typically involves the temporary storage of information essential for responding in working memory and the integration of the information with necessary knowledge, skills, and abilities stored in long-term memory – a process which is requisite to generating a response. These interactions between the test-taker and the test are referred to as *cognitive interactions*.

Further, most tests (including Test A) require the test-taker to respond (e.g., to complete a bubble sheet or click an answer choice); these tests require *physical interactions* between the test-taker and the test. Such response processes typically demand certain prerequisite knowledge, skills, and/or abilities. In addition, they require the application of the test-taker's perceptive, receptive, and cognitive abilities.

Another type of interactions is referred to as *emotive*. Emotive interactions involve any feelings that may be influenced by test features, or that may moderate other interactions. Motivation, typically defined as the desire to perform, is one of these emotions. Motivation interacts with the test insofar as the test-taker must possess the desire to interact with the test before he or she deliberately participates in the test event. Test self-efficacy may influence a test-taker's motivation for testing. Anxiety produces another kind of emotive interaction. Whether the test taker is worried about the potential consequences of the test, or whether his or her worries involve other issues, anxiety can interact with the test in a number of ways. Anxiety, like distraction, can draw needed resources from essential tasks as the test-taker focuses on nonessential stimuli, either real or imagined. For example, the siphoning of cognitive resources from processing requisite stimuli may affect the test-taker's cognitive processing (i.e., moderate the integration between working and long-term memory), potentially limiting his or her ability to meet the cognitive demands of the test. To the extent negative emotive interactions can be eliminated from the test event, the higher the likelihood the result will be reflective of the test-taker's optimal performance.

It should be noted that up to this point, there has been no discussion of any test event interaction that explicitly involves the target construct of Test A. Indeed, the test has generated an event comprised of perceptive, receptive, cognitive interactions, and perhaps emotive

interactions. Further, the test has required the participant to demonstrate certain knowledge, skills, or abilities simply to generate a random response. This would be true even if the participant possessed none of the targeted knowledge, skills, or abilities our hypothetical test was designed to measure.

Collectively, these other interactions are referred to as *ancillary interactions*. To the extent a test event fails to account for the ancillary interactions, the test may actually be measuring them. It is essential, therefore, that a test developer identify the individual characteristics a test-taker may contribute to the test event. Moreover, it is incumbent on the test developer to ensure that, to the extent possible, any and all potential interactions between the test-taker and the test are not permitted to impinge on the test result or any subsequent inferences. To yield a score that reflects only the target construct, the test must be designed to account for all of these other interactions. Unless the test user can be confident that the sum of these ancillary interactions has no effect on the test outcome, the accessibility of the test event should be considered as less than optimal.

When the effect of the aggregate ancillary interactions on the test event is null, what remains, of course is the interaction between the test and all of the intended domains of knowledge, skills, and/or abilities targeted for measurement by the test as possessed by the test-taker. This is referred to as the *targeted interaction* and is the only interaction from which the test should ultimately be designed to gather information. In practice, of course, the target interaction itself does not yield an actual *amount* of the target construct. Rather, the test event yields information (e.g., a test score) from which an inference can be made according to a priori guidelines. This inference should represent, in essence, a direct statement about the amount of the construct possessed by the test-taker.

A perfect test, therefore, measures the amount of the target construct possessed by the test-taker – nothing more, and nothing less. It may account for ARCs, but it does not measure them. We can thus conclude with an operational definition of accessibility situated within the framework of accessibility theory, as follows: Given that Subject J possesses X amount of Construct C, Test A is an optimally accessible measure of Construct C if the sum of interactions between Test A and Subject J is equal to X. To the degree that the sum of interactions between Test A and Subject J deviates from X, the test is not optimally accessible. In essence, the sum of the interactive effects in the test event on the test outcome should equal the effect of the targeted interaction. When ARCs influence the test outcome and impinge on subsequent inferences, there is an accessibility problem.

Given the goal of establishing a test event that yields a score from which inferences about the test-taker's amount of the target construct are valid, accessibility theory provides a useful foundation not only for establishing test events that contain as little contamination from ARCs as possible, but also for understanding accessibility across the educational milieu. Conceptually, this milieu contains three agents with the potential to effect a change in the outcome of the test: the school, the test, and the learner/test-taker. Figure 1 represents a unified model of access (UMA) in which arrows are used to indicate the causal relations among these and other variables.

Optimal test accessibility is manifest when the test event presents no barriers that preclude the test-taker from demonstrating the extent to which he or she has learned the tested content. Test access involves interactions among physical, perceptive, receptive, emotive, and/or cognitive factors that affect the test-taker's ability to respond to the demands of the test. One test-taker may experience barriers that preclude him or her from demonstrating his or her knowledge of the target construct of a test due to characteristics unique to that individual,

whereas another test-taker with different characteristics may enjoy unfettered access to the test. Thus, if the goal is to increase test scores, the test-taker clearly is a putative source of that change (after all, the test is designed to measure the amount of the target construct possessed by the test-taker.) For some students, testing accommodations (i.e., changes to the test administration or response features) may be used to address identified needs such that the test event yields a result that is uncontaminated by access-related interactions, and from which subsequent inferences reflect purer measures of the target construct.

Curricular access is manifest when the student is given the opportunity to learn the tested content. Opportunity to learn is a necessary condition for inferential validity of test scores. To the extent the student is not afforded the opportunity to learn the material, test score inferences about his or her achievement in the classroom may be invalid. These inferences may in turn result in decisions based on false conclusions. A *valid* inference about a student who has not been offered instruction in the test content is that the low test score reflects the test-taker's lack of opportunity to learn the material. By contrast, an invalid inference is that the student received instruction and sufficient opportunity to learn the material and failed to achieve at the expected level for some other reason. Thus, it is conceptually plausible to intervene at the school (or classroom) level if evidence suggests that an access problem exists here. Instructional methods, curricula, school resources, and myriad other variables influence opportunity to learn. Moreover, it should be noted that, like test accessibility, access between the school and the test-taker is bidirectional, and many of the same factors that influence access to testing also play a determinative part in the learning process (i.e., in the test-taker's role as a student.) Namely, unless the test-taker/student possesses all of the attributes and abilities necessary to overcome

any potential barriers to either his or her learning or the measurement of his or her learning, some or the entire access burden rests on some other change agent in the access model.

Insofar as it can be agreed, therefore, that there exist demands for learning or the measurement of learning that are in any degree imposed upon the test-taker, the burden for ensuring the student has optimal access both to learn the tested content and to demonstrate his or her extent of mastery of this content rests in some degree on the education and assessment systems. Even where evidence suggests one or more of the barriers precluding optimal access is within the control of the individual (e.g., motivation or effort), the instruction and assessment systems must be designed to account for these potential limiters.

Thus, it is essential for test developers to account for these limiters in the design process to ensure, to the extent possible, the accessibility of tests without the need for individualized accommodations for students with special needs. Based on the earlier discussion of accessibility theory for testing, three specific areas of study can be useful for pursuing this goal: cognitive load theory (CLT; Chandler & Sweller, 1991), research on test and item development, and guidance on computer and web accessibility.

Cognitive Load Theory. In his now famous "Magical Number Seven, Plus or Minus Two" article, Miller (1956) presented a synthesis of the research on what he termed *channel capacity* – i.e., the amount of information a person is able to process about a given stimuli, also known as *working memory*. Across a series of studies investigating participants' channel capacity for several variables including auditory pitch and loudness, taste, and visual identification of size and position, Miller reported the mean channel capacity was approximately seven categories (i.e., number of discriminable pitches or loudnesses, concentrations of saltwater, and object sizes or position, respectively). Across variables, the standard deviation was

approximately 3 with an overall range of 3 to 15 categories. Channel capacity was slightly higher when participants were permitted to identify categories on the basis of two or more variables (e.g., saltiness and sweetness for taste, pitch and loudness for audio stimuli, position and size for visuals, hue and saturation for color). Miller was surprised, however, at the minimal degree to which multidimensionality appeared to augment participants' capacity for processing information.

Cognitive load theory is a logical and theoretical extension of Miller's (1956) work. Until CLT was applied to assessment (e.g., in the research described in the following section), the theory singularly used as a model for understanding the demands of learning tasks and is grounded in the assumption that the mind has a limited capacity (i.e., in working memory) for processing information. In essence, CLT proponents posit that to properly gain knowledge from instruction, students must: (a) attend to the presented material, (b) mentally organize the material into a coherent structure, and (c) integrate the material with existing knowledge. Thus, the efficiency of instructional tasks depends on the extent to which the cognitive resources needed for this process are minimized.

Accordingly, CLT disaggregates the cognitive demands of learning tasks into three load types*: intrinsic load*, *germane load*, and *extraneous load*. Intrinsic load refers to the amount of mental processing that is requisite for completing a task. Germane load refers to cognitive demands that are not necessary for gaining essential knowledge but enhance learning by facilitating generalization or automation (e.g., lessons that require learners to extend learned concepts to arenas outside the classroom or apply them to novel situations). Extraneous load refers to the demand for cognitive resources to attend to and integrate nonessential elements that are preliminary to actual learning, but are nonetheless required for a learning task. Proponents of

CLT argue that learning tasks should be designed with the goal of minimizing the demand for cognitive resources that are extrinsic to the goals of instruction. The triune model of cognitive load was encapsulated by Paas, Renkl, and Sweller (2003): "Intrinsic, extraneous, and germane cognitive loads are additive in that, together, the total load cannot exceed the working memory resources available if learning is to occur"(p.2).

Intrinsic load contains all essential elements for understanding a task. The intrinsic load for simple tasks may require a small number of elements that may be understood apart from one another; more complex tasks may require understanding of, and interaction among, several elements. Paas, Renkl, and Sweller (2003) provided the example of learning the assignments of the set of 12 function keys on a typical QWERTY computer keyboard. Each element (i.e., an individual function key) may be understood apart from any other. By contrast, learning how to edit a photo on a computer requires several elements (e.g., changing color tones, darkness, contrast), all of which must be understood interactively to complete the task. The demands on working memory imposed by the intrinsic load of high-complexity learning tasks are greater than those imposed by simpler tasks. Decreasing the intrinsic load of a learning task results in a simpler task.

Based on Miller's (1956) assertion that working memory is an inherent human limitation, learning tasks with greater intrinsic load may not only require the learner to memorize the essential elements of the task, but also to integrate them. If a person with the capacity for storing three elements in his or her working memory is presented with a task requiring the interaction of 10 or more elements, the learner must combine sets of elements in a logical manner to permit the recall of these elements for the purpose of interaction. To do so, the learner must be able to store categorical combinations of multiple informational elements, or *schemas*, in his or her long-term

memory for later use. To the extent that the learner must utilize organizational schemas to gain knowledge from instruction, the amount of intrinsic load required for the task is increased. Thus, when the goal of a learning task is to produce more sophisticated understanding, an increase in the intrinsic load of the task may be inevitable.

Germane load refers to the cognitive demands of a learning task that may result in generalization of knowledge beyond that which results from the intrinsic elements of a task but which are not required for gaining initial or essential knowledge. Germane cognitive load enhances learning by targeting the development of schemas and automation. Consider the earlier example of learning computer keyboard function key assignments. If the instructor designed the instruction such that following the simple memorization of the function keys, the task concluded with students executing a series of function-key commands in a number of different software applications, the addition of this subsequent generalization activity would add germane load to the learning task.

Extraneous load, when required by a learning task, is preliminary to (or concurrent with) attending to the task, organizing the material into an existing structure, and integrating the material with existing knowledge. While germane load enhances learning, extraneous load interferes with learning by demanding the use of working memory for elements that are not essential for learning the material.

CLT primarily has been used to generate findings from which to provide direct instructional implications, specifically with regard to the adequacy of particular instructional designs. Chandler and Sweller (1991) described a series of studies conducted in Australia on electrical engineering trade apprentices. The results of these experiments indicated that cognitive load appeared to be lower when essential information disaggregated across two or more sources

was integrated (e.g., textual statements describing a diagram were embedded in the diagram itself). Based on lower test scores and longer processing time for learners who were given the "split-source" diagrams, the authors concluded that "presentation techniques frequently result in high levels of extraneous cognitive load that influence the degree to which learning can be facilitated….For this reason…examples that require learners to mentally integrate multiple sources of information are ineffective"(Chandler & Sweller, p.295). As such, the predominant implications for instructional practice pertained to the integration of graphics and visual representations with corresponding textual concomitants to reduce extraneous load.

Well into the second decade after the inception of CLT, the now-apparent application of cognitive load to multimedia learning began to emerge. Chandler and Sweller (1996) defined multimedia instruction as the presentation of words and images or other media to foster learning. Chandler and Sweller described two negative effects that may result from the improper structuring of multimedia instruction. The first is the *split attention effect*, whereby unintegrated split-source information in the presentation of material forces the learner to integrate the information to learn. The authors did not recommend integrating dual-source information in every instance, however. When one source of information contains all that is necessary to convey the material, the authors suggested the other source of information should be eliminated entirely to prevent the *redundancy effect*, whereby learners are distracted and bogged-down with excessive material.

Much of the recent CLT work has advanced these early applications of the theory to inform the development of newer multimedia instruction. For instance, based on Mayer and Moreno's cognitive theory of multimedia learning, the primary receptive senses – the ears and eyes – serve different functions when presented with a multimedia stimulus such as an

instructional lesson that includes both words and images. The ears hear the words, while the eyes perceive both the words and the corresponding images. In working memory, the sounds and images are organized into verbal and pictorial models, respectively. Finally, the models are integrated with prior knowledge and stored in long-term memory.

Mayer and Moreno (2003) argued the potential is high in multimedia learning for "cognitive overload"(p.43) and provided five scenarios in which cognitive overload may occur, as well as research-based guidelines for preventing them. The authors employ three novel concepts to describe these scenarios: *essential processing, incidental processing, and representational holding*. Essential processing basically corresponds to intrinsic load and refers to the cognitive demand required to make sense of presented material (i.e., selecting, organizing, and integrating words and images). Incidental processing corresponds to extraneous load and refers to the demand from nonessential aspects of the instructional material. Representational holding refers to the demand required to retain verbal or visual information in working memory.

Notwithstanding the broad overlap between instruction and testing, CLT heretofore has had little research application to school-age students with or without special needs or to the assessment of student learning. Considering the numerous similarities between instructional tasks and the variety of tasks required in many forms of tests, information about the cognitive load demands that may impact a test-taker's ability to demonstrate performance on assessments can be used to inform test development.

To the extent the cognitive demands of an assessment are intrinsic to the target constructs of the assessment, inferences made from test results are likely to represent the person's actual competence on the constructs. Extraneous load demands by an assessment item interferes with the test-taker's capacity to respond (i.e., demonstrate performance on the target construct) and

should be eliminated from the assessment process. Further, germane load, while enhancing learning at the instructional level, should be considered for elimination as well: unless an assessment task has the dual purpose of both instruction and assessment, the items on a test should demand only those cognitive resources intrinsic to the target constructs they are intended to measure. Indeed, the addition of germane load to an assessment task may represent an increase in the depth of knowledge of an item if it requires additional elements or interactivity among elements. Thus, the decision to include or exclude germane load from assessment tasks should be made deliberately.

Clark, Nguyen and Sweller (2006) synthesized the CLT research and generated a set of 29 guidelines for maximizing efficiency in learning. The majority of the recommendations focus on reducing redundancy, eliminating nonessential information from text and visuals, and integrating information from dual sources. There are also a number of cautionary considerations when using audio to supplement instruction.

Despite its tangential connection to CLT work, research on *interestingness* – the inclusion of details in text with the purpose of engaging the reader – applies directly to the design of accessible tests. For nearly 100 years, educators have debated the inclusion of nonessential text in reading material. Some advocates have recommended adding some nonessential text to increase interest, while adversaries have pointed to evidence that its inclusion may have a deleterious effect on comprehension. While a full understanding has not been reached, there is considerable research to support the use of caution when considering the inclusion of nonessential material in assessment tasks. At a minimum, the issue warrants a brief review.

In 1913, Dewey admonished educators to avoid attempting to improve educational lessons by including nonessential content with the sole intent of increasing interest. Seventy years later, Graves et al. (1988) conducted a study known as the "Time-Life Study" that became the polestar in a series of studies by several research groups over a decade that investigated the extent to which non-essential text, included only for the purpose of increasing interestingness, increased or decreased reader recall of main ideas. In the Time-Life Study, Graves et al. asked three groups of editors to revise history texts to improve their interest level. Of the three resulting texts, the revision characterized by the addition of low-importance, high-interest text (seductive details) were: (a) rated highest in interest level by readers and (b) recalled to a greater degree than other edits. Subsequent research, however, disconfirmed the hypothesis that seductive details increase recall (e.g., Garner, Alexander, Gillingham, Kulikowich, & Brown, 1991; Britton, Vandusen, Gulgoz, and Glynn, 1989; and a replication of the initial Time-Life study by Graves et al., 1991). Indeed, Garner, Gillingham, and White (1989) found that the addition of seductive details resulted in decreased recall. A subsequent investigation indicated texts that included seductive details took longer to read (Wade, Schraw, Buxton, & Hayes, 1993); thus, the authors theorized that seductive details draw attention away from main ideas to the detriment of reader recall. A review of these and other investigations of what has been termed the *seductive detail effect*, however, raised methodological questions and the authors called for further research (Goetz & Sadoski, 1995).

More recently, Harp and Mayer (1998) conducted a set of four experiments that provided confirmatory evidence of the seductive detail effect. The authors theorized that seductive details, rather than distracting or disrupting readers, prime inappropriate schemas around which readers then attempt to organize information for later recall. Schraw (1998) found both context-

dependent and context-independent seductive details were recalled better than main ideas, but only texts that included context-dependent seductive details took longer to read and results showed no significant effect of seductive details on reader recall of main ideas. Thus, while Schraw concluded the additional elaborative processing required to comprehend seductive details neither enhanced nor hindered recall, his results confirmed the earlier finding that seductive details increased reading load.

Thus, while scientific consensus has not been reached with regard to the extent to which test developers should include test features with the purpose of increasing motivation and interest, based on the body of research on the seductive detail effect and consistent with the recommendations of Mayer and Moreno (2003) and others, test developers should be careful to distinguish between nonessential information included with the intent of increasing interestingness and nonessential text that may interfere with the target construct and add extraneous cognitive load.

Research on test and item development. It is essential that the application of accessibility theory draws upon the collective expertise of test and item development scholars. More than a quarter-century prior to the inception of legislation permitting the AA-MAS for proficiency reporting for students with IEPs, Beattie, Grise, and Algozzine (1982; 1983) conducted experimental work on several test design features, many of which subsequently have been integrated in the majority of current large-scale assessments. For instance, Beattie et al. used format changes including the use of unjustified text for reading comprehension passages, placing passages in shaded boxes to set them apart from other text, including examples at the beginning of each new item section, adding arrow and stop-sign icons to the corners of test pages, and including response bubbles in the test booklet rather than using a separate answer sheet. Results

across two studies of students identified with a learning disability ($N = 345$ students in grade 3 and $N = 350$ students in grade 5) indicated the modifications increased students' scores without altering the target construct of the test.

Haladyna, Downing, and Rodriguez's taxonomy of recommendations for writing multiple-choice items is included in Figure 3. The taxonomy was based on a comprehensive review of research, as well as Haladyna's (1999) text on constructing and validating multiple-choice items (the latter of which is in the process of being updated.) Of particular relevance in this taxonomy are the various guidances on writing answer choices (i.e., the *key*, or correct response, and the *distractors*, or incorrect responses.) Further, based on a meta-analysis of over 80 years of research on item development, Rodriguez (2005) concluded that three answer choices are optimal for multiple-choice items. The author indicated that reducing items from 4 or 5 answer choices to 3 tends to result in nonsignificant or positive effects on the discriminatory power of items, nonsignificant changes in item difficulty, increased reliability of scores and, ultimately, a positive effect on the subsequent validity of inferences from results. As applied to the development or modification of tests with a focus on accessibility, Rodriguez' conclusion suggests best practice is to reduce the number of response options of multiple-choice items to three when it is feasible to do so.

Content concerns

1. Every item should reflect specific content and a single specific mental behavior, as called for in test specifications (two-way grid, test blueprint).
2. Base each item on important content to learn; avoid trivial content.
3. Use novel material to test higher level learning. Paraphrase textbook language or language used during instruction when used in a test item to avoid testing for simply recall.
4. Keep the content of each item independent from content of other items on the test.
5. Avoid over specific and over general content when writing MC items.
6. Avoid opinion-based items.
7. Avoid trick items.
8. Keep vocabulary simple for the group of students being tested.

Formatting concerns

9. Use the question, completion, and best answer versions of the conventional MC, the alternate choice, true-false (TF), multiple true-false (MTF), matching, and the context-dependent item and item set formats, but AVOID the complex MC (Type K) format.
10. Format the item vertically instead of horizontally.

Style concerns

11. Edit and proof items.
12. Use correct grammar, punctuation, capitalization, and spelling.
13. Minimize the amount of reading in each item.

Writing the stem

14. Ensure that the directions in the stem are very clear.
15. Include the central idea in the stem instead of the choices.
16. Avoid window dressing (excessive verbiage).
17. Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface.

Writing the choices

18. Develop as many effective choices as you can, but research suggests three is adequate.
19. Make sure that only one of these choices is the right answer.
20. Vary the location of the right answer according to the number of choices.
21. Place choices in logical or numerical order.
22. Keep choices independent; choices should not be overlapping.
23. Keep choices homogeneous in content and grammatical structure.
24. Keep the length of choices about equal.
25. *None-of-the-above* should be used carefully.
26. Avoid *All-of-the-above*.
27. Phrase choices positively; avoid negatives such as NOT.
28. Avoid giving clues to the right answer, such as
    a. Specific determiners including always, never, completely, and absolutely.
    b. Clang associations, choices identical to or resembling words in the stem.
    c. Grammatical inconsistencies that cue the test-taker to the correct choice.
    d. Conspicuous correct choice.
    e. Pairs or triplets of options that clue the test-taker to the correct choice.
    f. Blatantly absurd, ridiculous options.
29. Make all distractors plausible.
30. Use typical errors of students to write your distractors.
31. Use humor if it is compatible with the teacher and the learning environment.

(Haladyna, Downing, & Rodriguez, 2002)

Figure 3. *A taxonomy of multiple-choice item writing guidelines.*

Computer-based testing and web accessibility. In Bennett's (2001) "How the Internet Will Make Large-Scale Assessment Reinvent Itself," the author argued the advancing pervasiveness of computer-technology into all areas of modern life would lead, for better or worse, to the inevitable subsumption of standardized testing. Indeed, the ever-increasing use of online testing across the range of student assessment types supports his hypothesis. Commensurate with this apparent trend, it behooves developers of online tests to ensure adequate attention is paid to the accessibility of these tests for as many students as possible.

In addition to integrating the aforementioned CLT research on multimedia instruction, the Web Accessibility Initiative (W3; 2008a) has published guidelines to ensure web content is accessible to all users, including those with disabilities. According to W3, the Web Content Accessibility Guidelines (WCAG) were adopted as a web standard in 2009. The guidelines focus on four key principles: *perceivable*, *operable*, *understandable*, and *robust*. First, content is perceivable if it is presented to users in a way that is visible to at least one of their senses. Second, interface components and navigation are operable if the user is able to operate the interface. Third, the user must be able to understand both the content and the interface. Finally, robustness refers to content that permit reliable interpretations "by a wide variety of user agents, including assistive technologies"(W3, 2008b).

Test Accessibility Research

Recently, assessment researchers have begun to focus on accessibility as an issue critical to the validity of tests and test items. Still, there has been little focus on empirically documenting the accessibility of tests and test items. There have been several efforts made at providing guidance for test developers to address design issues that may hinder test-taker access (e.g.,

NCEO, 2002, 2006; Thurlow et al., 2009). Researchers at Vanderbilt University developed the Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, & Elliott, 2008) and TAMI Accessibility Rating Matrix (ARM; Beddow, Elliott, & Kettler, 2009b) to evaluate tests and test items with a focus on their accessibility for the intended test-taker population.

Commensurate with the development of the TAMI, several federally-funded studies have yielded empirical data informing the development of accessible tests and contributing to an increased understanding of test accessibility. These studies have emerged from two projects aimed at supporting states' efforts to develop an AA-MAS for students with persistent academic difficulties. The current study represents a critical extension of this line of research on test accessibility.

Consortium for Alternate Assessment Validity and Experimental Studies (CAAVES). In anticipation of aforementioned regulations under NCLB permitting states to develop alternate assessments based on modified achievement standards (AA-MASs), members of the six-state CAAVES project (Elliott & Compton, 2006-2009), in collaboration with Discovery Education Assessment (DEA), conducted a series of studies to examine the differential effects of item modifications on the performance of students who likely would be eligible for these tests and those who would not be eligible. Using a large item pool provided by DEA, an item modification team consisting of members of the consortium including assessment experts, educators, and educational psychology professors, modified 39 reading items and 39 mathematics items with the goal of enhancing their accessibility for students who would be eligible for an AA-MAS.

Assessment leaders in four of the participant states implemented computer-based field tests to examine the differential impact of the item modifications. The participant sample was disaggregated into three groups: (a) students without IEPs who would be ineligible for an AA-

MAS, labeled as students without disabilities (SWOD), (b) students with IEPs who likely would be ineligible for an AA-MAS, labeled as students with disabilities, not eligible (SWD-NE), and (c) students with IEPs who likely would be eligible for an AA-MAS, labeled as students with disabilities, eligible (SWD-E; $N = 694$ in reading, $N = 709$ in mathematics). Participants were randomly assigned to one of several forms of the tests, which were segmented into three counterbalanced parts: (a) original items, (b) modified items, and (c) modified items with reading support via audio-recorded narration of directions and ancillary text. Elliott et al. (2010) found the modified items were easier for all groups, and main effects for each group also were significant: the mean score for the SWOD group was significantly larger than the mean score for the SWD-NE group, and the SWD-NE group's mean score was significantly larger than the mean score of the SWD-E group. These results were consistent across both content areas. In reading, effect sizes for condition (i.e., original items vs. modified items vs. modified items with reading support) were large (partial $\eta^2 = .17$); in mathematics, the main effect size for condition was small (partial $\eta^2 = .05$). For reading, Cohen's *d* effect sizes for the modified condition over the original condition were moderate to large: $d = .50$ for the SWD-E group, $d = .49$ for the SWD-NE group, and $d = .38$ for the SWOD group. For mathematics, effect sizes for the modified condition over the original condition were smaller: $d = .31$ for the SWD-E group, $d = .25$ for the SWD-NE group, and $d = .20$ for the SWOD group. The effect sizes for the reading support condition over the modified condition were small across groups and content areas (Cohen's *d* range = .01 to .11).

Elliott et al. (2010) examined the practical implications of the modifications by applying cut-score proxies in the form of percentile ranks to the score distributions. Results indicated the effect of modification for the eligible group was comparatively greatest when the proficiency

criterion was low. Rasch model analyses indicated the percentage of students whose scores increased by more than one standard error of measure with modification was higher for the SWD-E group for both content area tests. Additionally, each of the field tests contained survey questions following the test items which asked students to reflect on specific modification strategies. Based on student response data from a post-test survey following the field test items, Roach, Beddow, Kurz, Kettler, and Elliott (2010) found students' overall perception of the modifications was positive.

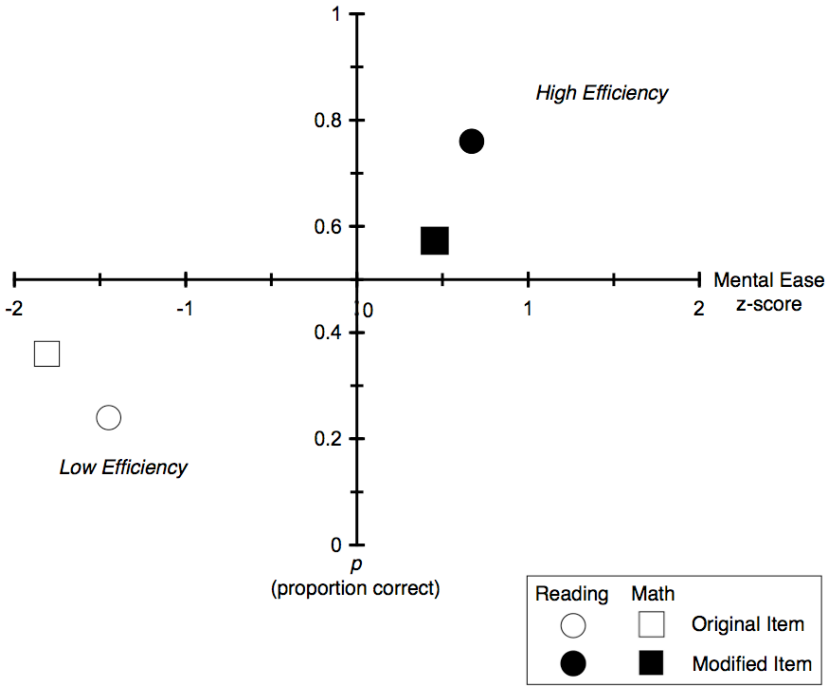
Kettler et al. (2008) delved further into the field test data by examining whether item difficulties within a Rasch model reflected an interaction paradigm whereby eligible students experienced a differential boost from modifications over the ineligible students. The authors found the decreases in item difficulty for the SWD-E group were significantly greater than for the SWOD or SWD-NE groups. Further, the authors found the internal consistency (Cronbach's alpha) coefficients were high and consistent across groups, conditions, and content areas.

Consortium for Modified Alternate Assessment Development and Implementation (CMAADI). The CMAADI project (Elliott, Roach, & Rodriguez, 2008-2010) is a collaboration with the departments of education of Arizona and Indiana in pursuit of a set of objectives to develop an operational alternate assessment based on modified achievement standards (AA-MAS). The initial work was with Arizona and involved item writing teams consisting of teachers, assessment developers, and assessment researchers with the purpose of evaluating and modifying a large pool of reading and mathematics items across multiple grade-levels. To prepare item modification teams for the modification session, the project leaders provided a half-day training session on universal design principles, cognitive load theory, and item and test development research. Additionally, teams were trained in the use of the TAMI (Beddow,

Kettler, & Elliott, 2008). Following training, teams divided into grade-level teams and modified items.

At the conclusion of the session, personnel at the Arizona Department of Education's Office of Assessment and Accountability revised the items according to the recommendations of modification teams and compiled dual sets of brief test forms consisting of a combination of original and modified items, with their sibling items contained in each alternate test form. Using these forms, researchers from Vanderbilt University conducted a cognitive interview ("think-aloud") study with a sample of 42 students in grades 4-8 and high school. Specifically, students were randomly assigned to one of the two test booklets for their respective grade-level and asked to complete the test items in the booklet while reporting their thoughts about the items. Following each section of the test, students were asked to indicate the relative difficulty of each item for them on a scale of 1-7. These self-reported ratings were transformed into Cognitive Ease *z*-scores and plotted against the student's performance on each item to determine the relative cognitive efficiency of the items prior to, and following, modification for accessibility. Additionally, each student's reading and mathematics teacher completed a Curricular Experiences Survey to indicate the extent to which the content measured by the test items had been covered in class during the school year, a proxy for the student's opportunity to learn the tested content. The results of this extensive cognitive interview study suggested item modifications generally produced changes in the item statistics in the expected direction as well as effected positive student perceptions as evidenced by their self-reported cognitive ease. Based on the small samples in this study, it was not apropos to conduct inferential analyses to confirm these observations. To illustrate the salient characteristics of the resulting data, the researchers adapted a visual representation of cognitive efficiency used by Clark and colleagues to

demonstrate the relations between modification and other dependent item variables. Figure 4 consists of a cognitive efficiency data plot for two multiple-choice items in original and modified form, with difficulty on the Y axis and student-reported cognitive ease (a proxy for cognitive load) on the X axis. By plotting items in this way, it is possible to observe the change in item characteristics as a function of modification. As described by Elliott, Kettler, Beddow, and Kurz, the "high efficiency" items are plotted in quadrant 1, which indicates the items are low in difficulty and low cognitive load. Low efficiency items, by contrast, are plotted in quadrant 3, indicating the items have high difficulty and high cognitive load. Thus, items that are successfully modified to enhance their accessibility are expected to move in an upward-right direction, progressing from quadrant 3 toward quadrant 1.



(Elliott, Kettler, Beddow, & Kurz, 2009)

Figure 4. *Cognitive efficiency plot*

In early 2009, a subset of the modified items in grades 7 and 10 from the initial modification session underwent subsequent revisions commensurate with the team's data-based conclusions from the cognitive interview study. These items were assembled into test booklets. In April of 2009, students across the state of Arizona participated in the statewide AIMS (Arizona Instrument to Measure Standards) assessment, which included the modified items in their original forms. In May of 2009, 300 students in grades 7 and 10 with and without IEPs completed the aforementioned brief test booklets containing modified versions of a subset of the items they received a few weeks before the general AIMS test. Additionally, a subsample of these students completed Maze-CBM reading fluency probes. Results of an internal analysis of these data indicated items generally functioned as expected, but some items had poorly-functioning distractors (i.e., answer choices with positive point-biserial correlations.) This indicated some modifications may have reduced the discrimination of their respective items.

Additionally, CMAADI personnel conducted focus groups with students following each the May pilot test sessions. Focus groups consisted of small homogenous groups of students with and without IEPs (i.e., those identified with disabilities, those not identified with disabilities) to collect their observations, perceptions, and thoughts about tests, test items, and about the particular modifications that were made to the test items they received during the study. These focus groups were audio-recorded for later analysis. Initial analyses of these data indicated students were interested in engaging the testing process to show what they know, and they perceived modification strategies as favorable. Several students in these focus groups indicated they did not feel they had received sufficient instruction in the tested content and were anxious about the potential consequences of their presumably poor performances on the state test.

Concurrently, a team of CMAADI investigators used the Accessibility Rating Matrix (ARM; Beddow et al., 2009b) to examine the accessibility of a large pool of multiple-choice and constructed-response items in Indiana across grades 3-8 in Language Arts, Mathematics, Science, and Social Studies (N = 166 items). All of the items had undergone one round of modifications by the Indiana assessment team. The accessibility review team consisted of four educational psychology professors and three graduate students in assessment. Item accessibility was rated on the 4-point ARM scale (4 = Maximally accessible for nearly all test-takers; 1 = Inaccessible for many test-takers). Ratings were disaggregated by item elements (i.e., passage or item stimulus, item stem, visuals, answer choices, and page/item layout.) As per the ARM rating procedure, raters also assigned an overall accessibility rating to each item. Results indicated the item pool consisted of items with a broad range of accessibility levels. Across the item pool, the mean overall item accessibility was high. The evaluation team identified several positive attributes of the reviewed set of items and identified a number of modifications to improve the accessibility of the items. Detailed feedback was provided for each item reviewed. Results of reliability analyses conducted following the completion of the item review indicated reliability on ARM Overall Accessibility ratings was high; specifically, pairs of expert raters agreed within 1 level (i.e., perfect or adjacent agreement) for 87% of items. Predetermined reliability procedures specified that for items where perfect or adjacent agreement was not reached, the item was rated by a third expert rater. Thus, a third expert conducted accessibility ratings for 13% of items.

Research Questions and Predictions

This study was motivated to advance understanding of accessibility as it influences learning and testing. It specifically was designed to address three research questions:

Question #1. What are the effects of testing accommodations, item modifications, and a combination of the two on test performance for students with different abilities? Research and theory by Sireci et al. (2005) and Elliott et al. (2009) led to the prediction that testing accommodations and item modifications would improve test performance; as well, the combinative impact of testing accommodations and modifications would be significantly greater than either strategy used alone. Additionally, Hollenbeck (2002) suggested access is a continuous variable, and the current study disaggregates accessibility methods into three levels. Based on the differential boost observed in previous accommodations and modifications work, I expected students with the highest level of need to have greater score increases for all three experimental conditions compared to students with a lower level of need.

Question #2. What are the relations between item accessibility and other psychometric indices used to characterize items? Based on Elliott et al.'s (2009) research the relation between item accessibility and item difficulty would be moderate. Based on research by Rodriguez (2005) and Haladyna et al. (2002), the magnitude of relations between accessibility and item discrimination would be low.

Question #3. How do students with different abilities perceive the accessibility of items, their cognitive demand, their teachers' coverage of the content, and their own predicted performance? The guiding research by Roach et al. (2009) suggested when compared to student perceptions of original items, students perceptions of modified items would be positive. Although this question is exploratory, results of the CMAADI cognitive interview study (Roach,

2009) suggested students would report lower perceived difficulty, lower cognitive load, better understanding, and better predicted performance when asked about modified items compared to original items. Adolescent students with and without IEPs were expected to report accommodations as only moderately helpful and moderately likeable.

Each of these fundamental questions and related predictions is directly testable. The following chapters describe the method and results of a two-phase study that provides data-based evidence for testing these questions.

CHAPTER II

METHOD

The proposed study consisted of two phases: an Experimental Phase and a Follow-up Questionnaire Phase. The first phase used a 2 x 4 experimental design with two groups of seventh-grade students: (a) students with an IEP (i.e., receiving Special Education services; n = 103), and (b) students with no IEP (n = 329; see Figure 5). Participants from both groups were randomly assigned to take a 34-item math test in one of four conditions: Form A (original items), no accommodations (control condition AN); (b) Form A, accommodations (experimental condition AA); (c) Form B (items that have been modified to improve their accessibility), no accommodations (experimental condition BN); and (d) Form B, accommodations (experimental condition BA).

Participants

Participants were students in grade 7 (N = 432) from two middle schools in California and four middle schools in Arizona. Table 1 contains student participant demographic data including IEP status, sex, ethnicity, disability category for the students with IEPs, and English Language Learner (ELL) status. Recruitment efforts focused on a target sample of 320 (n = 160 with no IEP; n = 160 with an IEP) in anticipation of an effect size of .50 for accommodations and modifications, based on existing research on accommodations and modifications.[2] It is generally expected the large majority of students who will be eligible for an AA-MAS will be identified with a learning disability in reading or mathematics. Of students with IEPs (n = 103),

75% (n = 77) were eligible for special education services under the identification of Specific Learning Disability in either reading or mathematics. Of the 71% of the total sample for which state assessment data were available for the past two years, 47% of students with IEPs (n = 26) had scored below proficient on the reading or mathematics domain on the state assessment for the past two years, compared to 13% of students with no IEP. Seventh-grade mathematics teachers (N = 17) also participated in the study for the purpose of selecting student accommodations. Each participating school (N = 6) received a $1200 honorarium.

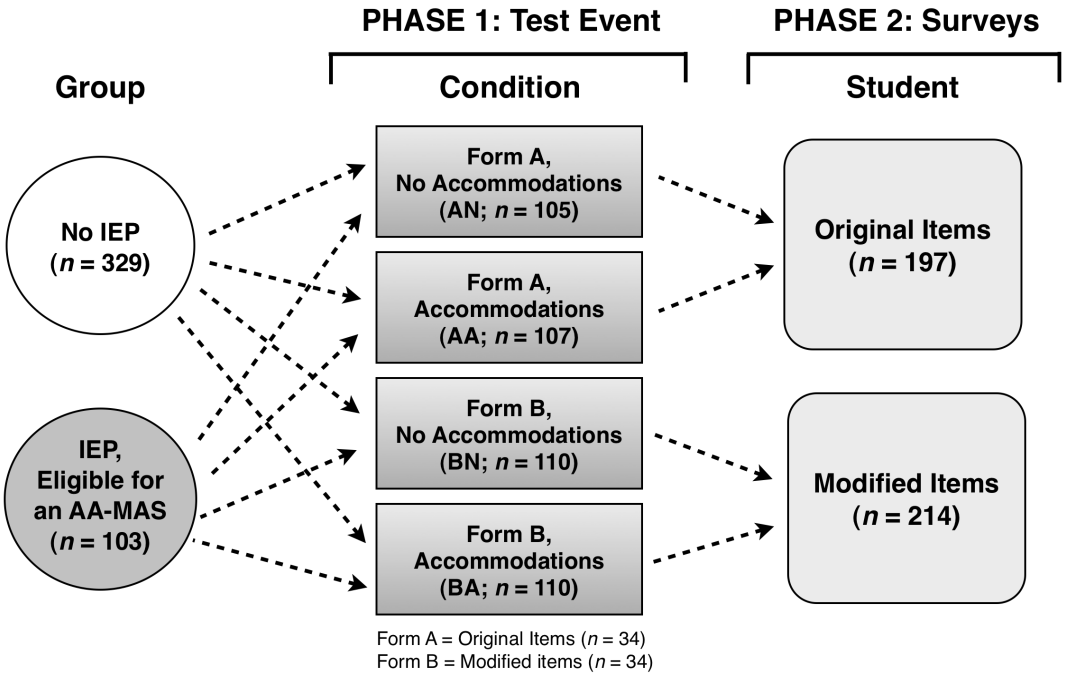


Figure 5. *Study design.*

Table 1.
Participant Demographics

	Original Items		Modified Items	
	No Accomm. *n* (%)	Accomm. *n* (%)	No Accomm. *n* (%)	Accomm. *n* (%)
Eligibility Status				
No IEP (*n* = 329)	78 (74%)	82 (77%)	82 (75%)	87 (79%)
IEP (*n* = 103)	27 (26%)	25 (23%)	28 (25%)	23 (21%)
Sex				
Female	47 (45%)	56 (52%)	64 (58%)	47 (43%)
Male	58 (55%)	51 (48%)	46 (42%)	63 (57%)
Ethnicity*				
Asian	2 (4%)	0 (0%)	1 (1%)	5 (8%)
Black	6 (11%)	2 (3%)	2 (3%)	3 (5%)
Caucasian	40 (71%)	47 (77%)	48 (72%)	45 (73%)
Hispanic	7 (13%)	11 (18%)	14 (21%)	9 (15%)
Other	1 (2%)	1 (2%)	2 (3%)	0 (0%)
Disability Category	(*n* = 27)	(*n* = 25)	(*n* = 28)	(*n* = 23)
Specific Learning Disability	18 (67%)	17 (68%)	24 (86%)	18 (78%)
Autism	2 (7%)	1 (4%)	1 (4%)	0 (0%)
Speech/Language Impairment	2 (7%)	1 (4%)	1 (4%)	0 (0%)
Emotional Disturbance	2 (7%)	1 (4%)	0 (0%)	2 (9%)
Other Health Impairment/ 504	3 (11%)	5 (20%)	2 (7%)	3 (13%)
English Language Learners*				
Yes	0 (0%)	1 (2%)	0 (0%)	1 (2%)
No	56 (100%)	60 (98%)	67 (100%)	61 (98%)
Total	105 (24%)	107 (25%)	110 (25%)	110 (25%)

Note. Ethnicity and ELL data were not available for all participants.

Materials

Test forms. Participants were randomly assigned to complete one of two test forms developed for the current study. Form A consisted of 34 mathematics items from the Discovery Education Assessment grade 6 test item bank. Grade 6 items were used because the study was conducted in the Fall of the participants' seventh-grade year, so they had not had sufficient opportunity to learn grade 7 content. Items on Form A were delivered in their original form, having not undergone modification procedures to address accessibility concerns. Form B consisted of 34 of the same items from Form A, in modified form (i.e., item siblings) using the Accessibility Rating Matrix (Beddow et al., 2009). To control for opportunity-to-learn, the investigator conducted an alignment analysis to ensure that all items mapped to state standards from both states for either grade 5 or grade 6. The results of this analysis are presented in Table 2. Based on the proportions of items for each content strand, alignment and coverage are similar for both Test Form A and Test Form B.

Accessibility Rating Matrix (ARM). The ARM is a decision-making tool for evaluating and modifying tests and test items with a focus on enhancing their accessibility for all test-takers. For the current study, the investigator and a trained team of item raters with expertise in assessment and students identified with disabilities used the ARM to rate the accessibility of the original items on a criterion-based scale.

Table 2.
Comparison Between AIMS and STAR Blueprints and Study Test Forms.

Strand / Concept	Arizona Instrument to Measure Standards (AIMS) # of items measuring (%)	California Standardized Testing and Reporting (STAR) # of items measuring (%)	Test Forms (A and B) # of items measuring (%)
1. Numbers and Operations	17 (25%)	25 (39%)	7 (21%)
2. Data Analysis / Probability / Discrete Numbers	13 (19%)	11 (17%)	11 (32%)
3. Patterns / Algebra / Functions	13 (19%)	19 (29%)	9 (26%)
4. Geometry and Measurement	15 (22%)	10 (15%)	5 (15%)
5. Structure and Logic	10 (15%)	Embedded	2 (7%)
Across Standards	68 items	65 items	34 items

Note. Some items measure multiple content strands. Thus, the total number of items listed across content strands is greater than the number of items on the test.

The ARM is divided into two matrices: Item Analysis and Overall Analysis (see Figures A1-A4). The ARM is used to evaluate the accessibility of the item according to the basic elements of a multiple-choice test item: item stimulus, item stem, visuals, answer choices, and page or item layout, as well as to indicate any revisions or modifications that are likely to improve the accessibility of the item. Levels of accessibility on the ARM are based on the extent to which the item is *maximally accessible* for a given portion of the intended test-taker population. A maximally accessible item is an item that contains no barriers that would limit or hinder the test-taker from demonstrating his or her knowledge of the target construct. An item that is maximally accessible for nearly all test-takers, therefore, requires few, if any, cognitive

resources in excess of those needed to show what the test-taker knows. If the extraneous cognitive load demand of an item differentially impacts performance on the item across the test-taker population, the item's accessibility rating is less than optimal. In essence, a maximally accessible item should be accessible to 95-99% of the test-taker population, thus reducing threats to test score validity from incomplete access for fewer than 5% of test-takers.

Results of several validity studies indicate the content of the TAMI and TAMI ARM is valid for the purpose of measuring test item accessibility and that expert raters can be trained to score item accessibility with a high degree of reliability (Beddow, Kettler, & Elliott, 2010). The TAMI Technical Supplement (Beddow et al., 2009a) contains a thorough description of validity research on the TAMI and ARM. Evidence based on this research indicates that the TAMI ARM is useful for evaluating item accessibility and is sensitive to differences in accessibility across a range of items.

Prior to the test event, an independent TAMI evaluation team consisting of expert reviewers including three assessment professors and two graduate students used the TAMI ARM (Beddow et al., 2009b; see Figures A1-A4) to apply accessibility review procedures to each of the items in both test forms. Modification procedures for Test form B are described in the next section.

Student follow-up questionnaire. Following each test, student participants received one of nine versions of a questionnaire about some of the items they saw on the test, each containing a different set of three items to be rated, for a total of 18 items (9 original, 9 modified) across the sample. For each test item, students were asked to respond to a set of question, as follows: (a) "How well did you understand this question?" (b) "How hard did you have to work to answer this question?" and (c) "How much have you been taught about this in school?" and (d) "How

sure are you that you got this question right?" For students in the accommodated condition in which they were able to ask for reading support during the test, the survey contained the following question: "Did the adult help you to do better on the test?"

Procedures

Item and test form development. Original items for Form A consisted of items from the Discovery Education Assessment (DEA) Grade 6 item bank. Grade 6 items were used because the study took place during October of the participants' seventh-grade year. To ensure all students had theoretical access to the content knowledge of each of the items on the test, Form A consisted of a set of 34 grade 6 multiple-choice mathematics items that matched at least one content standard from California and Arizona's standards for grade 6 (see Table 2.)

When Test Form A was compiled, the primary investigator, two assessment professors from Vanderbilt University, one quantitative methods professor from the University of Minnesota, and one advanced graduate student in educational psychology evaluated all of the items according to the principles of accessibility theory and suggested changes to improve the accessibility of the items. Modifications included reducing answer choices from four to three, simplifying language in item stimuli and item stems, and reducing the complexity of visuals (for further detail about the suggested modifications, see the TAMI Accessibility Rating Matrix; Beddow et al., 2009). The investigator then modified items according to these suggestions. The format of the modified items differed from the standard DEA delivery format, in which items are coded using the web language HTML, using tags for formatting changes and embedded images where needed. For the modified test form, the investigator developed individual item images based on the suggestions of the rater team; each image contained the item stimulus, stem, and

visuals. Additional images contained each answer choice. Following each set of modifications, the evaluators reviewed the items to determine whether further changes were necessary to improve the accessibility of the items. Reliability of ratings was optimized according to the agreement procedures established for all prior reviews: specifically, for items for which Rater 1 and Rater 2 did not have perfect or adjacent agreement for the ARM Overall Accessibility rating, the item was discussed among the evaluation team until consensus on a final item rating was reached. Following each rating, any item that did not receive a "4" rating overall and for all item elements was further modified and re-checked by the evaluation team. Due to limitations with the DEA item delivery system, it was not possible to incorporate all of the suggested item changes (e.g., increased font size, additional white space) while keeping each entire item on a single, visible screen. This resulted in layout concerns that precluded many of the items from receiving the highest accessibility rating. The modification process resulted in two similar forms of the test: Test A consisted of items that were unmodified and not revised prior to use in the study; Test B consisted of all of the items from Test A after modifications according to the principles described above.

Table 3 contains descriptive item accessibility data for each form, disaggregated by item element and overall. The accessibility ratings for the items, therefore, reflect the decreased accessibility of the modified items due to the need for test-takers to scroll down to see one or more of the answer choices. As a result of the fact that many of the items required the test-taker to scroll down to see one or more answer choices, few of the modified items received the highest accessibility rating. Overall, the mean overall accessibility rating of the modified form ($M = 2.9$, $SD = 0.5$) was higher than the original ($M = 2.7$, $SD = 0.5$; $t = 2.49$; $df = 66$; $p < .01$). Accessibility ratings for the item stimulus, item stem, visuals, and answer choices were higher

for Form B compared to Form A (all *p*s < .01). Due to the fact that the computer-delivery system for the items could not accommodate the demand to present the entirety of each item on a single viewable screen, the mean accessibility rating for the layout was higher for Form A ($p < .01$).

Participant recruitment and group assignment. Participants came from six schools: four were located in a large school district in Arizona, and two were located in a small district in Southern California. Districts were selected on the basis of their size and representativeness of the general population of the state. When permission was obtained at the district level, the investigator contacted the principals at each of the six schools and provided information about the study. When principals consented to participate, they sent consent letters to potential teacher participants with the request that willing teachers send consent letters home with their students with and without IEPs. Students who returned signed consent letters were included in the study.

Using stratified random assignment by IEP category (i.e., IEP versus no IEP), students were placed into one of four conditions: Test form A (original items) with no accommodations (AN; control group; $n = 105$), Test form A with accommodations (AA; $n = 110$); Test form B (modified items) with no accommodations (BN; $n = 110$); or Test form B with accommodations, (BA; $n = 109$). The randomization process was conducted on the day of the test event, immediately prior to student testing. Specifically, the investigator labeled each computer with a colored card that indicated one of the four conditions. When student participants entered the room, the investigator instructed them to sit at any computer of their choosing. For the majority of the study period, equal numbers of computers were labeled with each condition. For the final three days of the study, the numbers of cards were adjusted as needed to ensure approximately equal numbers of students were assigned to each condition.

Table 3.
Accessibility Ratings by Item Element and Overall

	Form A *M (SD)*	Form B *M (SD)*
Overall	2.7 (0.5)	3.2 (0.7)
	(*n* = 34)	(*n* = 34)
Stimulus	2.8 (0.6)	3.8 (0.5)
	(*n* = 32)	(*n* = 34)
Stem	2.9 (0.7)	4.0 (0.0)
	(*n* = 34)	(*n* = 34)
Visuals	2.8 (0.9)	3.8 (0.5)
	(*n* = 15)	(*n* = 21)
Answer Choices	3.0 (0.4)	3.5 (0.7)
	(*n* = 34)	(*n* = 34)
Page/Item Layout	3.6 (0.7)	3.5 (0.5)
	(*n* = 34)	(*n* = 34)

Note. Accessibility was rated on a 4-point scale (1 = Inaccessible for many test-takers; 2 = Maximally accessible for some test-takers; 3 = Maximally accessible for most test-takers; 4 = Maximally accessible for nearly all test-takers).

Selection of testing accommodations. Due to the larger number of student participants from Arizona compared to California, the Arizona accommodations were used for both states. Table 4 contains selection frequencies and percentages of accommodations by student group. For students with an IEP, the checklist instructed teachers to select accommodations based on what is listed on the IEP. For students with no IEP, teachers assigned accommodations based on their instructional knowledge of what may benefit the students on the test. Teachers completed accommodations checklists prior to the test event.

Table 4.
Accommodations Selected by Group and Availability to Participants

Accommodation	No IEP	IEP	How Implemented
	n (%)	*n* (%)	
Testing in a small group;	144 (42%)	85 (92%)	Standard.
Testing one-on-one;	88 (26%)	17 (18%)	Available upon request.
Testing in a separate location or in a study carrel;	72 (21%)[1]	45 (49%)	Available upon request.
Being seated in a specific location within the testing room or being seated at special furniture;	24 (7%)	41 (45%)	Available upon request.
Having the test administered by a familiar test administrator;	190 (56%)	35 (38%)	Not available.
Using a special pencil or pencil grip;	0 (0%)	1 (1%)	Available upon request.
Using devices that allow the student to see the test: glasses, contacts, magnification, etc.	18 (5%)	3 (3%)[2]	Available regardless of study condition.
Having questions about the directions students read on their own answered;	60 (18%)[1]	33 (36%)	Available upon request.
Place marker use[1];	49 (14%)[1]	11 (12%)[2]	Available upon request.
More breaks and/or several shorter sessions[1];	77 (23%)[1]	35 (38%)	Available upon request.
Test at a different time of the day;	20 (6%)	7 (8%)	Available upon request.
Simplify scripted directions;	163 (48%)	58 (63%)	N/A – Test had no scripted directions.
Read aloud or sign the directions students read on their own[1];	51 (15%)[1]	62 (67%)	Available upon request.
Read aloud in English or sign the test items;	101 (30%)[1]	61 (66%)[2]	Available upon request.
Have answers transferred from the test book into an answer document;	2 (1%)	2 (2%)	Precluded by the use of computer-delivery,
Record or dictate multiple-choice responses to a scribe;	0 (0%)	3 (3%)	Available upon request.
Use of a personal whiteboard[1]	69 (20%)[1]	10 (11%)	All participants received paper/pencil.

[1]Selection frequency was significantly lower for students in the accommodated condition compared to the non-accommodated condition ($p < .05$).
[2]Selection frequency was significantly higher for students in the accommodated condition compared to the non-accommodated condition ($p < .05$).

During the test event, reading support (i.e., read-aloud of the directions and test questions) was available to all students in accommodated conditions (AA and BA.) To ensure the integrity of any additional selected accommodation during the test event, the research personnel attached the accommodations checklists to parent consent letters and consulted them during each test event to ensure any additional selected accommodations were available to the students in the accommodated condition. During the test event, the research team tracked the use of on-demand accommodations by coding them on individual student surveys.

Test event. Each school's test events took place in the school's computer lab. All of the six school computer labs were arranged in a similar fashion, which consisted of 20-30 computer carrels that included a keyboard, mouse, screen, and tower. Each workspace permitted sufficient space to facilitate students' use of provided scratch paper during the test. Test event proctors included the primary investigator and 1-2 trained research assistants. Assistants consisted of one school principal with 10 years of teaching experience, one assessment professional, and one advanced graduate student in educational leadership.

Prior to the test events at each school, school principals collected signed consent forms from each participating teacher, and collected signed parent consent forms for each potential student study participant. Each potential participant's teacher completed an Accommodations Checklist for the potential student participants. The principal and/or a designated staff member at the school then generated a schedule whereby each participant would be sent by their mathematics teacher to the school computer lab to participate in the study. The investigator received the schedule, participant rosters, consent forms, and checklists from the principal or designated staff member prior to delivering the tests to study participants.

Prior to each test period, the investigator placed colored index cards at each computer carrel. Each color corresponded to one of the four study conditions: NO (orange), NM (yellow), AO (green), and AM (pink). The investigator logged into the Discovery Education Assessment website on each computer and selected the test form that corresponded to the condition indicated by the card. To ensure random assignment to conditions, the investigator instructed students to sit at any computer upon entering the room. Once seated, the investigator verbally delivered the following script to students:

> *"Thank you for coming today. Your parents have agreed to let you be part of a research project. The project is about testing. Raise your hand if you took the [Arizona or California] state tests last Spring in school. I think you took a bunch of tests in reading, and a bunch in math, maybe some in social studies or science. This research project is about making tests that are more fair for students, so you can show what you know and can do. You are going to be taking a 34-question math test today. Your score will not be reported to your teacher. Your score will not be reported to your parents. It will not be reported to your princial, or your school. I will use your score to help me to learn about how to make tests that are better for students. The questions on the test are 6th-grade math questions. Why are they 6th-grade questions? Because I didn't think it would be fair to give you 7th-grade questions, because you have only been in 7th-grade for a few months, so you haven't had the chance to learn all of the 7th-grade material. This is not a timed test. You will have all the time you need to finish all 34 questions. After you finish the test, I will give you a short survey. The survey asks you to look at 3 of the questions you just saw on the test. You will answer a few questions about each of them: first, How well did you understand this question? Very well, not very well, or somewhere in*

between. Circle one of the numbers, whichever one you want. The next question asks, How hard did you have to work to answer this question? Very hard, not very hard, or somewhere in between. The next question asks you, How much have you been taught about this in school: A lot, or very little, or somewhere in between. Maybe your teacher covers this stuff every day, and you'll circle 6, or maybe you say, "I've never seen this stuff before in my life," and you'll circle 1. Now, if you are sitting at a computer with a PINK or GREEN card, you may get help on the test. You may ask for the questions to be read aloud to you. We can't help you answer the question, but we can read it to you, or help you read a word, or we can even sit with you and read every question to you. All you need to do is raise your hand and ask for help. If you are sitting at an ORANGE or YELLOW card, you just do the test as you normally would. We want to see whether the kids who get help actually do better on the test, or if it doesn't make any difference."

Once all the students had initiated the test, the investigator announced once again, *"Remember, if you are at a computer with a PINK card, or a GREEN card, you may ask for help on the test. We can read the questions aloud to you, or we can help you with one word. We can even sit with you during the test and read every question out loud to you. Since this test is not a reading test, it doesn't matter if we read it to you, because the test isn't a test of how well you can read. Some students like hearing the questions read aloud; they feel it helps them do better on the test. If you want one of us to read one of the questions aloud to you, or a word, or even the whole test, just raise your hand and ask. Again, if you're sitting at a computer with a PINK or GREEN card, you can ask for help reading the questions."*

Additionally, if a student's accommodations checklists indicated some other type of accommodation that was not already addressed by the standard test conditions, the investigator

or a research consultant approached that student to deliver the accommodation, or to let him or her know that the accommodation was available to him or her. For the two Accommodations conditions (i.e., AA and BA), the investigator informed students that accommodations were available, but did not deliver reading support unless students requested it. Of the students in the two conditions for which accommodations were available (n = 217), only 15 students requested them (7%).

Student follow-up questionnaire. During each test period, the investigator or a research assistant gave each student a post-test survey corresponding to his or her assigned experimental condition. The surveys consisted of 3 test items with a set of questions for each item (as described above; see Appendix C.) The investigator or an assistant explained the survey to each student individually once he or she was finished with the computer-based math test. When students were finished, they returned to class.

Data Analyses, Expected Outcomes and Criteria for Evaluating Outcomes

The principal goals of the study were to examine the differential effects of a range of strategies for increasing access to mathematics test items. To wit, the investigator conducted analyses with the specific purpose of examining the hypothesized relations among several accessibility variables. The following predictions follow from the stated research questions:

Prediction #1. There are expected to be positive main effects for each of the experimental conditions against the control group. Specifically,results of unpaired t tests will indicate significantly greater total scores for each of the AA, BN, and BA conditions compared to the AN condition (p_{crit} = .05 / 3 = all ps < .017). For all three experimental conditions, students in the IEP group were expected to experience a significantly greater "boost" in total scores compared to

students with no IEPs, as evidenced by a significant interaction between group and condition using a 2-way ANOVA ($p < .05$), as well as greater Cohen's *d* effect sizes for the IEP group compared to the No-IEP group.

Prediction #2. The association between item accessibility and key psychometric indices was expected to be positive and to a moderate degree. Specifically, the relations between item accessibility as rated by the TAMI ARM and difficulty (*p*) was expected to result in a moderate (between .30 and .50) correlation, while the correlation between ARM rating and item discrimination (i.e., point-biserial correlations) was expected to be low (between .10 and .20). Both correlations, however, were expected to be of a magnitude that would be considered statistically significant ($p < .05$).

Prediction #3. Students' comprehension, cognitive ease, and perceived performance were expected to be significantly higher for students who received modified items compared to students who received original (unmodified) items. Specifically, student ratings of comprehension, cognitive ease, and perceived performance for the combined modified-accommodated (BA) and modified-nonaccommodated (BN) conditions were expected to be significantly higher than ratings for the combined original nonaccommodated (AN) and original accommodated (AA) conditions ($p < .05$) based on results of unpaired *t* tests. In addition, comprehension, cognitive ease, and perceived performance were expected to be lower for students identified with disabilities than for their non-identified peers. Finally, ratings of comprehension, cognitive ease, and comprehension for students in the IEP group were expected to be significantly lower than ratings for students with no IEPs ($p < .05$) based on results of unpaired *t* tests. Further, students who received accommodations, helpfulness of the accommodations is expected to be higher for students identified with disabilities than for their

non-identified peers. That is, for students in the AA and BA conditions, mean helpfulness and desirability of accommodations is expected to be significantly higher for students with IEPs than for students with no IEPs ($p < .05$), based on results of unpaired *t* tests.

CHAPTER III

RESULTS

The study was conducted as planned over the course of two weeks in October. The basic data to address the research questions and related predictions are portrayed in Table 4.

Test Forms

Table 5 contains psychometric statistics for each test form, disaggregated by group when possible. Each test form consisted of 34 grade 6 mathematics items that were aligned to the Arizona and California state mathematics content standards for grade 5 or grade 6. Standard (z) scores were generated for each participant. For Test Form A, the mean item difficulty (i.e., the proportion of students who responded correctly to each item) was 0.56 (range = .30 to .80). For Form B, the mean item difficulty was 0.64 (range = .30 to .90). We calculated the readability of each item using five standard readability indices (as is common practice at assessment companies): The Flesch-Kincaid Grade Level, the Gunning-Fog Score, the Coleman-Liau Index, the SMOG Index, and the Automated Readability Index. For Form A, the mean item readability based on these five indices, was approximately grade 6 (M = 5.9, SD = 2.5). For Form B, mean item readability was at grade 4 (M = 4.0, SD = 2.3). The difference was significant, $t(66) = -3.27$, $p < .01$. Not including words contained in graphs and charts, Form A contained 33% more words than Form B (1276 words in Form A compared to 852 words in Form B). Mean item word counts for Forms A and B were 37 and 25, respectively. The difference in word count was significant, $t(66) = -2.65$, $p < .05$. From Form A to Form B, the number of words in each item

was reduced by a mean of 26%. Appendix B contains each test item in original and modified form, with corresponding item accessibility, difficulty, discrimination, word count, and readability statistics.

Table 5.
Psychometric Statistics by Form and Participant Group

		Form A (Original) *M* (*SD*)	Form B (Modified) *M* (*SD*)
Total Score	Total	18.99 (6.49)	20.35 (4.80)
		(*n* = 212)	(*n* = 220)
	No IEP	20.61 (6.02)	21.68 (4.15)
		(*n* = 160)	(*n* = 169)
	IEP	13.98 (5.20)	15.94 (4.17)
		(*n* = 52)	(*n* = 51)
Total Score (*z*)	Total	-0.11 (1.13)	0.12 (0.83)
	No IEP	0.17 (1.05)	0.36 (0.72)
	IEP	-0.98 (0.90)	-0.64 (0.72)
Cronbach's α	Total	0.84	0.79
	No IEP	0.82	0.72
	IEP	0.77	0.70
SEM	Total	0.37	0.36
	No IEP	0.35	0.38
	IEP	0.37	0.38
Difficulty (*p*)	Total	0.56	0.64
	No IEP	0.61	0.69
	IEP	0.41	0.49
Readability		5.90 (2.55)	3.98 (2.27)
Word Count		37.52 (21.85)	25.06 (16.66)

Cronbach's α coefficients for each test form were within the acceptable range (Form A, α = .84; Form B, α = .79). For students in the No-IEP group, α was .82 for Form A and .72 for Form B. For students in the IEP group, α was .77 for Form A and .70 for Form B. Using these coefficient alphas as reliability estimates for each test form, I calculated the standard error of measurement using the following formula, where s_E is the standard error of measurement, s_X is the standard deviation of the test form, and r_{xx} is its reliability estimate:

$$s_E = s_x \sqrt{1 - r_{xx}}$$

Using each respective form's overall SD and reliability coefficient, the SEM for Form A was 0.37; the SEM for Form B was 0.36. To calculate SEM by group, I used the SD and reliability coefficient for the form for each respective group. For the No-IEP group, the SEM was 0.35 for Form A and 0.38 for Form B. For the IEP group, the SEM was 0.37 for Form A and 0.38 for Form B.

Student Performance

Descriptive statistics disaggregated by experimental condition and group (Table 6) indicated overlapping score distributions across experimental conditions; further, contrary to Prediction 1, the mean test score for students in the AN condition (control) was higher than the mean test score for students in the AA condition. A 2x4 group-by-condition analysis of variance (ANOVA), however, was significant, $F(7, 424) = 18.48$, $p < .01$. There was not a significant between-group difference for experimental condition, $F(2, 428) = 2.43$, $p < .07$. The ANOVA indicated a significant within-group difference, $F(1, 424) = 118.24$, $p < .01$. There was no interaction between condition and group, $F(3, 424) = 0.83$, $p = .48$. To detect the source of the difference between groups, pairwise comparisons were made. To correct for multiple between-

groups comparisons, alpha was divided by 4 (p_{crit} = .05 / 4 = .0125), using Form A with no accommodations (AN) as the control condition. The difference in total score between students in the accommodated condition on Form A was not significant, $t(210) = -0.75$, $p = .77$. The difference between the BN group and the control group was not significant, $t(213) = 1.02$, $p = .15$. The difference between the BA group and the control group also was not significant, $t(213) = 1.63$, $p = .05$. The difference between the two groups who participated in Test B (modified items) and the two groups who participated in Test A (original items), however, was significant, $t(431) = 2.49$, $p < .01$.

Table 6.
Student Performance, Descriptive Statistics, 2 x 4 Design

	Test A (Original)				Test B (Modified)			
	No Accommodations (AN)		Accommodations (AA)		No Accommodations (BN)		Accommodations (BA)	
	No IEP (*n* = 78) *M (SD)*	IEP (*n* = 27) *M* (*SD*)	No IEP (*n* = 82) *M (SD)*	IEP (*n* = 25) *M* (*SD*)	No IEP (*n* = 82) *M (SD)*	IEP (*n* = 28) *M* (*SD*)	No IEP (*n* = 87) *M (SD)*	IEP (*n* = 23) *M* (*SD*)
Difficulty (*p*)	.62	.41	.58	.44	.63	.48	.64	.44
Mean	.57 (*n* = 105)		.55 (*n* = 107)		.59 (*n* = 110)		.61 (*n* = 110)	
Total score	21.2 (5.8)	14.0 (5.3)	20.1 (6.3)	14.0 (5.2)	21.3 (4.0)	16.5 (4.3)	22.0 (4.3)	15.3 (4.0)
Mean	19.32 (6.44)		18.65 (6.54)		20.1 (4.59)		20.6 (5.01)	
Total score (z)	0.3 (1.0)	-10 (0.9)	0.8 (1.1)	-1.0 (0.9)	0.3 (0.7)	-0.5 (0.8)	0.4 (0.7)	-0.8 (0.7)
Mean	-0.05 (1.12)		-0.17 (1.14)		0.08 (0.80)		0.17 (0.87)	

Table 7 contains Cohen's *d* effect sizes for the test score differences between each of the experimental conditions, by group and for the total sample. The main effect of modification (i.e., test form) was .23. The effect size for the IEP group was nearly twice that of the No-IEP group (.41 compared to .21). These differences all were statistically significant (all *p*s < .05).

As indicated earlier, it is standard practice in Arizona for test administrators to deliver the majority of individualized accommodations (including reading support) upon request only. Of the 217 students for whom accommodations were available (*n* = 107 for Form A, *n* = 110 for Form B), only 15 students (7%) actually requested and thus received them, notwithstanding repeated reminders by the research team (approximately 6 per hour) that they were available. I therefore did not anticipate a large effect of accommodations on student scores. Indeed, there was no significant main effect of accommodations on student total score for either form, $t(210) = -0.75$ for Form A; $t(218) = 0.77$ for Form B, both *p*s > .20.

Table 7.
Cohen's d Effect Sizes by Group and Condition

Group	B-A	A-N	BA-AN	BN-AN	AA- AN	BA-AA	BA- BN
Total sample	.23[a]	-.01	.21	.14	-.12	.33[a]	.08
No IEP	.21[a]	-.03	.15	.06	-.25	.40[a]	.09
IEP	.41[a]	-.14	.26	.52[a]	.20	.06	-.26

A = Test form A; B = Test form B; AA = Test form A, accommodated condition; AN = Test form A, nonaccommodated condition; BA = Test Form B, accommodated condition; BN = Test form B, nonaccommodated condition. [a]Difference was statistically significant ($p < .05$)

I thus did not observe, and was unable to detect statistically, a main effect of accommodations on student scores. Based on the significant difference between the mean total score of the combined groups who received modified items (BA and BN) and the groups who received original items (AA and AN), I simplified the design of the study for analytic purposes. Namely, I collapsed the four experimental conditions into two, as follows: AN and AA became simply Form A (unmodified / original), and BN and BA became Form B (modified). For the purposes of remaining analyses, the study design became a 2 x 2 (two experimental conditions by two groups; see Table 8.) Results of a 2x2 ANOVA with IEP status and modification condition (original items versus modified items,) using total score as the dependent variable, was significant, $F(3,428) = 1072.43$, $p < .01$. Specifically, there was a significant main effect of IEP status, $F(1, 428) = 117.70$, $p < .01$, as well as a significant main effect of modification condition (original items versus modified items), $F(1, 428) = 7.05$, $p < .01$. There was no interaction between IEP status and modification condition, $F(1, 428) = 0.61$, $p = .43$.

For the total sample, the mean score for Form B was higher than the mean score for Form A ($\Delta z = .23$). This difference was significant, $t(431) = 2.49$, $p < .01$). The mean score for students in the No-IEP group who took the modified form was higher than the mean score for students in the No-IEP group who took the original form ($\Delta z = .19$). This difference was significant, $t(327) = 1.88$, $p < .05$. This represented less than 1 SEM difference in scores between the two conditions for students in the No-IEP group. The mean score for students in the IEP group who took the modified form was higher than the mean score for students in the IEP group who took the original form ($\Delta z = .34$). This difference was significant, $t(101) = 2.11$, $p < .05$. This represented a difference of slightly less than 1 SEM in scores between the two conditions for students in the IEP group. On the original form, mean scores for students in the No-IEP

group scores were higher than those for students in the IEP group ($\Delta z = 1.34$). This difference was significant, $t(210) = 7.12$, $p < .01$. This represented a difference of greater than 3 SEMs between the No-IEP group and the IEP group on the original form. On the modified form, the mean score for students in the No-IEP group was greater than the mean score for students in the IEP group ($\Delta z = 1.19$). This difference was significant, $t(218) = 8.65$, $p < .05$. This represented a difference of greater than 2 SEMs between students in the two groups on the modified form.

Table 8.
Student Performance, Descriptive Statistics, 2 x 2 Design

	Test A (Original)		Test B (Modified)	
	No IEP (n = 160) *M (SD)*	IEP (n = 52) *M* (*SD*)	No IEP (n = 169) *M (SD)*	IEP (n = 51) *M* (*SD*)
Difficulty (p)	.60	.42	.64	.46
Mean	.56		.64	
Total score	20.6 (6.0)	14.0 (5.2)	21.7 (4.1)	15.9 (4.2)
Mean	18.99 (6.49) (n = 212)		20.35 (4.80) (n = 220)	
Total score (z)	0.2 (1.0)	-1.0 (0.9)	0.4 (0.7)	-0.6 (0.7)
Mean	-0.11 (1.13)		0.12 (0.83)	

Item Statistics

Descriptive accessibility rating data for the test forms using the ARM (Beddow et al., 2009) were presented in Table 3 (see Chapter II). The mean overall accessibility for Form A was 2.7 ($SD = 0.5$) compared to 3.4 for Form B ($SD = 0.7$). This difference was statistically significant, $t(66) = 3.88$, $p < .01$. Four of the five mean item element ratings were similarly

higher for Form B; the exception was the Item Layout element. The mean difference in mean ARM rating for Item Stimulus was statistically significant, $t(66) = 7.38$, $p < .01$. The mean difference in mean ARM rating for Item Stem was statistically significant, $t(66) = 9.50$, $p < .01$. The mean difference in mean ARM rating for Visuals was statistically significant, $t(34) = 4.12$, $p < .01$. The mean difference in mean ARM rating for Answer Choices was statistically significant, $t(66) = 3.33$, $p < .01$. The observed difference between the Item Layout rating for Form A ($M = 3.62$, $SD = 0.70$) and Form B ($M = 3.53$, $SD = 0.51$) was not statistically significant, $t(66) = 0.60$, $p = .28$.

Item difficulty statistics (i.e., proportions of test-takers who responded correctly to each item) for all 68 items across the two forms are documented in Tables 9 and 10. Additionally, the table includes the difference between the difficulties for the total sample and by group. The mean difficulty for Form A was .56. The mean difficulty for Form B was .64. For the total sample, a greater proportion of participants responded correctly on the modified item for 25 of the 34 items. For the remaining nine items, a greater proportion of students responded correctly on the original item. The results were the same for the No-IEP group. For the IEP group, the proportion of students who responded correctly was higher for the modified item for 24 of the 34 items (i.e., one fewer item than the No-IEP group). There were nine items for which the increase in p was at lesat two times greater for the IEP group compared to the No-IEP group. By contrast, there were 10 items for which the increase in p was at least two times greater for the No-IEP group compared to the IEP group.

Item discrimination (D) statistics for all 68 items across the two forms, for the total sample and disaggregated by group are document in Tables 11 and 12. The discrimination for an item is found by subtracting the mean difficulty for the item for the portion of the tested

population whose total score was in the lower 27% of participants from the mean difficulty for the item for the portion of the tested population whose total score was in the upper 27% of participants. Ebel (1954) argued item discrimination of .20 is low, and above .40 is high. The mean item discrimination for Form A was 0.49 (*SD* = 0.14). The mean item discrimination for Form B was 0.40 (*SD* = 0.19). The mean difference in item discrimination was -0.09 (*SD* = 0.17). For students in both groups, *D* was slightly lower than the mean for the total sample for both Form A (*M* = 0.37, *SD* = 0.14 for the No-IEP group; *M* = 0.38, *SD* = 0.24 for the IEP group) and Form B (*M* = 0.33, *SD* = 0.18 for the No-IEP group; *M* = 0.35, *SD* = 0.19 for the IEP group). The mean difference between items across the two forms was -0.03 (*SD* = 0.08) for the No-IEP group and -0.03 (*SD* = 0.32) for the IEP group.

Table 9.
Item Difficulties by Form and Participant Group, Items 1-16

Item		Form A	Form B	Change	Item	Form A	Form B	Change
1	**Total**	**0.58**	**0.60**	**+0.02**	**9**	**0.43**	**0.62**	**+0.19**
	No IEP	0.63	0.66	+0.03		0.49	0.71	+0.22
	IEP	0.44	0.43	-0.01		0.25	0.33	+0.08
2	**Total**	**0.45**	**0.49**	**+0.04**	**10**	**0.80**	**0.90**	**+0.10**
	No IEP	0.45	0.51	+0.06		0.86	.093	+0.07
	IEP	0.46	0.43	-0.03		0.63	0.80	+0.17
3	**Total**	**0.51**	**0.42**	**-0.09**	**11**	**0.55**	**0.75**	**+0.20**
	No IEP	0.58	0.49	-0.09		0.63	0.79	+0.16
	IEP	0.33	0.18	-0.17		0.31	0.61	+0.30
4	**Total**	**0.76**	**0.85**	**+0.09**	**12**	**0.52**	**0.66**	**+0.14**
	No IEP	0.81	0.89	+0.08		0.58	0.72	+0.14
	IEP	0.62	0.75	+0.13		0.37	0.49	+0.12
5	**Total**	**0.56**	**0.57**	**+0.01**	**13**	**0.50**	**0.59**	**+0.09**
	No IEP	0.60	0.65	+0.05		0.59	0.63	+0.04
	IEP	0.44	0.31	-0.15		0.23	0.45	+0.22
6	**Total**	**0.32**	**0.42**	**+0.10**	**14**	**0.80**	**0.91**	**0.11**
	No IEP	0.33	0.41	+0.09		0.84	0.96	+0.12
	IEP	0.29	0.47	+0.18		0.67	0.73	+0.06
7	**Total**	**0.64**	**0.88**	**+0.24**	**15**	**0.55**	**0.47**	**-0.08**
	No IEP	0.72	0.91	+0.19		0.58	0.53	-0.05
	IEP	0.38	0.78	+0.40		0.46	0.27	-0.19
8	**Total**	**0.25**	**0.49**	**+0.24**	**16**	**0.77**	**0.96**	**+0.19**
	No IEP	0.30	0.51	+0.21		0.82	0.99	+0.17
	IEP	0.12	0.41	+0.29		0.62	0.86	+0.24

Table 10.
Item Difficulties by Condition and Participant Group, Items 17-34

Item		Form A	Form B	Change	Item	Form A	Form B	Change
17	**Total**	**0.58**	**0.45**	**-0.13**	**26**	**0.37**	**0.30**	**-0.07**
	No IEP	0.59	0.44	-0.15		0.44	0.27	-0.17
	IEP	0.52	0.45	-0.07		0.17	0.41	+0.24
18	**Total**	**0.74**	**0.70**	**-0.04**	**27**	**0.33**	**0.43**	**+0.10**
	No IEP	0.80	0.73	-0.07		0.34	0.45	+0.12
	IEP	0.56	0.59	+0.03		0.29	0.35	+0.06
19	**Total**	**0.47**	**0.64**	**+0.17**	**28**	**0.59**	**0.71**	**+0.12**
	No IEP	0.53	0.67	+0.14		0.65	0.79	+0.14
	IEP	0.29	0.51	+0.22		0.40	0.43	+0.03
20	**Total**	**0.59**	**0.89**	**0.30**	**29**	**0.66**	**0.64**	**-0.02**
	No IEP	0.64	0.93	+0.29		0.68	0.67	-0.01
	IEP	0.42	0.76	+0.34		0.60	0.51	-0.09
21	**Total**	**0.53**	**0.64**	**+0.11**	**30**	**0.51**	**0.75**	**+0.24**
	No IEP	0.59	0.73	+0.14		0.58	0.83	+0.25
	IEP	0.35	0.31	-0.04		0.33	0.49	+0.16
22	**Total**	**0.54**	**0.52**	**-0.02**	**31**	**0.75**	**0.62**	**-0.13**
	No IEP	0.62	0.57	-0.05		0.79	0.69	-0.10
	IEP	0.29	0.33	+0.04		0.60	0.41	-0.19
23	**Total**	**0.30**	**0.59**	**+0.29**	**32**	**0.43**	**0.61**	**+0.18**
	No IEP	0.29	0.64	+0.35		0.49	0.70	+0.21
	IEP	0.31	0.43	+0.12		0.27	0.31	+0.04
24	**Total**	**0.65**	**0.82**	**+0.17**	**33**	**0.69**	**0.72**	**0.03**
	No IEP	0.71	0.88	+0.17		0.73	0.75	+0.02
	IEP	0.44	0.63	+0.19		0.56	0.63	+0.10
25	**Total**	**0.58**	**0.48**	**-0.10**	**34**	**0.70**	**0.75**	**0.05**
	No IEP	0.63	0.53	-0.10		0.76	0.79	+0.03
	IEP	0.44	0.29	-0.15		0.54	0.63	+0.09

Table 11.
Item Discrimination by Form and Participant Group, Items 1-16

Item		Form A	Form B	Change	Item	Form A	Form B	Change
1	**Total**	**0.49**	**0.40**	**-0.09**	**9**	**0.59**	**0.47**	**-0.12**
	No IEP	0.46	0.48	+0.02		0.38	0.33	-0.05
	IEP	0.54	0	-0.54		0.15	0.67	+0.51
2	**Total**	**0.25**	**0.29**	**+0.04**	**10**	**0.43**	**0.24**	**-0.20**
	No IEP	0.35	0.36	+0.01		0.21	0.17	-0.04
	IEP	0.15	0.33	+0.18		0.54	0.33	-0.21
3	**Total**	**0.63**	**0.58**	**-0.04**	**11**	**0.44**	**0.31**	**-0.13**
	No IEP	0.51	0.52	+0.02		0.27	0.21	-0.06
	IEP	0.23	0.50	+0.27		0.31	0.75	+0.44
4	**Total**	**0.47**	**0.18**	**-0.29**	**12**	**0.42**	**0.24**	**-.19**
	No IEP	0.30	0.12	-0.18		0.23	0.14	-0.09
	IEP	0.69	0.42	+0.28		0.62	0.08	-0.53
5	**Total**	**0.55**	**0.62**	**0.07**	**13**	**0.68**	**0.64**	**-0.04**
	No IEP	0.56	0.55	-0.01		-0.68	0.69	+0.01
	IEP	0.38	0.08	-0.30		0.38	0.25	-0.13
6	**Total**	**0.18**	**0.18**	**0.00**	**14**	**0.32**	**0.27**	**-0.05**
	No IEP	0.33	0.26	-0.07		0.16	0.10	-0.07
	IEP	-0.38	0.08	+0.47		0.62	0.50	-0.12
7	**Total**	**0.57**	**0.31**	**-0.26**	**15**	**0.33**	**0.47**	**+0.15**
	No IEP	0.20	0.21	+0.01		0.27	0.33	+0.06
	IEP	0.38	0.33	-0.05		0.38	0.33	-0.05
8	**Total**	**0.61**	**0.33**	**-0.28**	**16**	**0.44**	**0.13**	**-0.31**
	No IEP	0.48	0.29	-0.20		0.16	0.02	-0.14
	IEP	0.38	0.08	-0.30		0.46	0.33	-0.13

Table 12.
Item Discrimination by Condition and Participant Group, Items 17-34

Item		Form A	Form B	Change	Item	Form A	Form B	Change
17	**Total**	**0.33**	**-0.20**	**-0.13**	**26**	**0.35**	**0.02**	**-0.33**
	No IEP	0.28	0.29	+0.01		0.16	0.05	-0.11
	IEP	0.62	0.33	-0.28		0.23	0.08	-0.15
18	**Total**	**0.51**	**0.22**	**-0.29**	**27**	**0.42**	**0.22**	**-0.21**
	No IEP	0.32	0.26	-0.06		0.35	0.19	-0.16
	IEP	0.69	0.08	-0.61		-0.08	0.33	+0.41
19	**Total**	**0.44**	**0.27**	**-0.17**	**28**	**0.72**	**0.64**	**-0.08**
	No IEP	0.23	0.26	+0.03		0.46	0.43	-0.03
	IEP	0.23	0.25	+0.02		0.62	0.67	+0.05
20	**Total**	**0.57**	**0.27**	**-0.30**	**29**	**0.55**	**0.65**	**+0.11**
	No IEP	0.37	0.19	-0.18		0.57	0.67	+0.10
	IEP	-0.54	0.33	-0.21		0.77	0.42	-0.35
21	**Total**	**0.61**	**0.69**	**+0.08**	**30**	**0.77**	**0.58**	**-0.19**
	No IEP	0.46	0.52	+0.07		0.47	0.40	-0.06
	IEP	0.46	0.42	-0.04		0.62	0.33	-0.28
22	**Total**	**0.72**	**0.55**	**-0.17**	**31**	**0.47**	**0.60**	**+0.13**
	No IEP	0.46	0.48	+0.02		0.50	0.03	+0.31
	IEP	0.62	0.25	-0.37		0.31	0.42	+0.11
23	**Total**	**0.22**	**0.65**	**+0.44**	**32**	**0.65**	**0.69**	**+0.05**
	No IEP	0.38	0.50	+0.12		0.57	0.55	-0.02
	IEP	0.15	0.67	+0.51		0.23	0.25	+0.02
24	**Total**	**0.40**	**0.29**	**-0.11**	**33**	**0.53**	**0.29**	**-0.24**
	No IEP	0.15	0.07	-0.07		0.30	0.24	-0.06
	IEP	0.31	0.42	+0.11		0.31	0.50	+0.19
25	**Total**	**0.57**	**0.60**	**+0.03**	**34**	**0.49**	**0.51**	**+0.02**
	No IEP	0.57	0.55	-0.02		0.43	0.43	+0.00
	IEP	0.23	0.42	+0.19		0.08	0.58	+0.51

The correlation matrix between accessibility ratings and item difficulty is presented as Table 13. This relation indicates the degree to which the accessibility of an item is related to student performance on the item. Correlations between ARM ratings and item difficulty were mixed. For Form A, the correlation between item accessibility and item difficulty was very small and negative (r = -.07). For the No-IEP group, the correlation was very small and negative (r = -.04). For the IEP group, the correlation was small and negative (r = -.15) (ns). For Form B, the correlation between item accessibility and item difficulty was moderate (r =.25) for the total sample. For the No-IEP group, the correlation was moderate (r =.30). For the IEP group, the correlation was very small (r = .03.) The prediction that item difficulty would be moderately correlated with item accessibility, therefore, was unsupported.

Table 13.
Pearson Correlations Between Accessibility and Difficulty

	Form A	Form B
Total	-.07	.25
No IEP	-.04	.30
IEP	-.15	.03

The correlation matrix for accessibility and item discrimination is presented in Table 14. Correlations between accessibility and item discrimination were small to moderate for both forms (r = .27 and .40 for Form A and Form B, respectively), indicating that the degree to which items discriminate between test-takers of high and low total scores (i.e., the degree to which student performance on the items depends on their overall performance on the test) was related to the accessibility of the items. Correlations for Form A were .23 and .07 for the No-IEP and

IEP groups, respectively. Corresponding correlations for Form B were .09 and .40. All were nonsignificant. The expectation that item discrimination would be positively correlated with item accessibility, therefore, was supported, but the prediction that the correlations would be statistically significant was not.

Table 14.
Pearson Correlations Between Accessibility and Item Discrimination

	Form A	Form B
Total	.25	.42
No IEP	.24	.33
IEP	.07	.42

The correlation matrix between the number of words per item and item difficulty is presented as Table 15. Correlations all were negative, indicating items become easier as the number of words decreases. For Form A, the correlation between word count and difficulty was -.20 for the total sample. For Form B, the correlation was -.32 for the total sample. The correlation between the difference in the number of words between the two sibling items and the difference in the difficulty between the two items was -.30.

Table 15.
Pearson Correlations Between Word Count and Difficulty

	Form A	Form B
Total	.-.20	-.32
No IEP	-.22	-.35
IEP	-.12	-.29

Post-Test Surveys

Descriptive statistics for the student post-test surveys are presented in Table 16. As stated earlier, each student received his or her survey immediately upon completing the computer-delivered test. The survey consisted of a series of 4-5 questions about each of three of items to which the student had responded on the test (see Figure C1). Question 1 asked students to rate how well they understood the question, using a 6-point Likert-type scale (1 = Not very well; 6 = Very well). The mean rating for Form B (the modified form) was 4.89 (SD = 1.00) compared to a mean of 4.55 (SD = 1.03) on Form A (the original form). This difference was statistically signfiicant, $t(409) = 3.46$, $p < .01$. This was consistent with the expectation that students would report higher comprehension for the modified items compared to the unmodified items.

For both test forms, the mean understanding rating for students in the No-IEP group was higher than for students in the IEP group. These differences both were statistically significant, $t(195) = 2.91$, $p < .01$ for Form A; $t(212) = 5.33$, $p < .01$ for Form B. This was consistent with the expectation that students identified with disabilities would report decreased comprehension of the items compared to their non-identified peers.

Question 2 asked students to rate the cognitive load of the question by proxy (i.e., "How hard did you have to work to answer this question?") on a Likert-type scale (1 = Not very hard; 6 = Very hard). Student-reported cognitive load was lower for Form B (M = 2.50, SD = 1.02) compared to Form A (M = 2.73, SD = 1.04). The difference was statistically significant, $t(409) = 2.20$, $p < .05$. This was consistent with the expectation that students would report higher cognitive demand for the unmodified items.

Table 16.
Student Post-Test Survey Data, Descriptive Statistics

	Test A (Original)		Test B (Modified)	
	No IEP	IEP	No IEP	IEP
How well did you understand this question?[1]	4.67 (0.92) (n = 147)	4.19 (1.25) (n = 50)	5.08 (0.86) (n = 165)	4.27 (1.15) (n = 49)
	4.55 (1.03) (n = 197)		4.89 (1.00) (n = 214)	
How hard did you have to work to answer this question?[2]	2.62 (0.91) (n = 147)	3.03 (1.30) (n = 50)	2.35 (0.97) (n = 165)	2.99 (1.05) (n = 49)
	2.73 (1.04) (n = 197)		2.50 (1.02) (n = 214)	
How much have you been taught about this in school?[3]	3.86 (1.11) (n = 147)	3.47 (1.38) (n = 50)	4.10 (1.09) (n = 165)	3.67 (1.08) (n = 49)
	3.76 (1.19) (n = 197)		4.00 (1.10) (n = 214)	
How sure are you that you got this question right?[4]	4.29 (1.12) (n = 147)	3.95 (1.35) (n = 50)	4.83 (1.02) (n = 165)	3.72 (1.36) (n = 49)
	4.20 (1.19) (n = 197)		4.57 (1.20) (n = 214)	
How much did the adult(s) help you to do better on the test?[5]	2.32 (1.66) (n = 33)	2.69 (1.93) (n = 12)	3.41 (2.06) (n = 41)	3.65 (1.74) (n = 16)
	2.42 (1.72) (n = 45)		3.48 (1.96) (n = 57)	

[1] 1 = Not very well; 6 = Very well.
[2] 1 = Not very hard; 6 = Very hard.
[3] 1= Very little; 6 = A lot.
[4] 1 = Not very sure; 6 = Very sure.
[5] 1 = Not very much; 6 = Very much.

For both test forms, the mean self-reported cognitive load rating for students in the No-IEP group was lower than for students in the IEP group. These differences both were statistically significant, $t(195) = -2.43$, $p < .01$ for Form A; $t(212) = -3.92$, $p < .01$ for Form B. This was

consistent with the expectation that students identified with disabilities would report higher cognitive demand than non-identified students.

Question 3 asked students to rate their opportunity to learn the material (OTL) required to respond to the question (i.e., "How much have you been taught about this in school?") on a 6-point Likert-type scale (1 = Very little; 6 = A lot). The mean rating for OTL was lower for Form A ($M = 3.76$, $SD = 1.19$) compared to Form B ($M = 4.00$, $SD = 1.10$). This difference was statistically significant, $t(409) = -2.12$, $p < .05$. For both test forms, the mean OTL rating for students in the No-IEP group was higher than for students in the IEP group. These differences both were statistically significant, $t(195) = 2.02$, $p < .05$ for Form A; $t(212) = 2.44$, $p < .01$ for Form B.

Question 4 asked students to rate their confidence on the item (i.e., "How sure are you that you got this question right?") on a 6-point Likert-type scale (1 = Not very sure; 6 = Very sure). The mean rating for confidence was lower for Form A ($M = 4.20$, $SD = 1.19$) compared to Form B ($M = 4.57$, $SD = 1.20$). This difference was statistically significant, $t(409) = -3.18$, $p < .01$. This was consistent with the expectation that students would report higher confidence on the modified items compared to unmodified items. Additionally, for both test forms, the mean confidence rating for students in the No-IEP group was higher than for students in the IEP group. These differences both were statistically significant, $t(195) = 1.77$, $p < .05$ for Form A; $t(212) = 6.16$, $p < .01$ for Form B. This was consistent with the expectation that students identified with disabilities would report lower confidence than their non-identified peers.

Students who were assigned to one of the two accommodated conditions (AA or BA) were asked whether the adult(s) in the room helped them to do better on the test. This will heretofore be referred to as the perceived accommodation effect (PAE). The mean rating for

PAE was lower for Form A ($M = 2.42$, $SD = 1.72$) compared to Form B ($M = 3.48$, $SD = 1.96$). This difference was statistically significant, $t(100) = -2.85$, $p < .01$. For both test forms, students in the IEP group rated the helpfulness of accommodations higher than students in the No-IEP group. These differences, however, both were statistically non-significant, $t(43) = 0.64$, $p = .26$ for Form A; $t(55) = 0.40$, $p = .35$ for Form B. The expectation that students identified with disabilities would report that accommodations were helpful to a greater degree than their non-identified peers, therefore, was unsupported.

CHAPTER IV

DISCUSSION

This study was conducted within an accessibility theory framework, which is based on two primary assumptions: (a) first, that there exist a set of interactions between individual test-taker characteristics and features of achievement tests, and (b) that for some test-takers, these interactions reduce the validity of inferences that can be made from test results. In essence, some students possess certain characteristics (e.g., working memory limitations) that, while they are assumed to be orthogonal to the target construct(s) of an achievement test, in actuality interact with the test in such a way that test results are negatively biased and subsequent score inferences do not accurately reflect the measurement of his or her knowledge of the target construct of the test. In the context of accessibility theory, a dominant cause of the undesired impact of test-taker characteristics on test scores is incomplete or reduced *access*. To wit, a test item that contains features that interact negatively with the characteristics of a portion of the test-taker population is referred to as inaccessible for some test-takers. Solutions to this problem, therefore, must aim to increase access for more test-takers.

The current study was among the first experimental investigations of the effects of both testing accommodations and item modifications, two strategies that have been used with the goal of increasing access to assessment for students with special needs. Based on prior research, I predicted the main effects of each of these strategies would be replicated, and I predicted the combined effect of accommodations and modifications would be greater than their separate effects. Additionally, I expected to find a differential boost in test scores for students identified

with disabilities compared to their non-identified peers, for both strategies alone, as well as in combination. Similarly, I predicted positive relations between the accessibility of the items and student performance, as well as between accessibility and item discrimination.

Research Questions

The principal objective of the current study was to examine the differential effects of two strategies for increasing access to mathematics test items. The analyses reported in the previous section were conducted to test three predictions related to this objective.

Question #1. What are the effects of testing accommodations, item modifications, and a combination of the two on test performance for students with different abilities?

Given the research design employed, it was expected there would to be positive main effects for each of the experimental conditions against the control group.Specifically, results of unpaired *t* tests will indicate significantly greater total scores for each of the AA, BN, and BA conditions compared to the AN condition. For all three experimental conditions, students in the IEP group were expected to experience a significantly greater "boost" in total scores compared to students with no IEPs, as evidenced by a significant interaction between group and condition using a 2-way ANOVA, as well as greater Cohen's *d* effect sizes for the IEP group compared to the No-IEP group. The prediction was partly supported. The results indicated there was no detectable effect of accommodations on student test performance for either group or for the total sample. There was, however, amoderate effect for modification (i.e., Form). Moreover, the magnitude of this effect for students with IEPs was nearly twice that of their general education peers. Results indicated there was no additive effect of modifications and accommodations. As indicated previously, few students availed themselves of available accommodations; as a result, I

collapsed the study design into two experimental conditions, based on item modifications: Form A (original) versus Form B (modified) for the balance of the analyses.

Question #2. What are the relations between item accessibility and other psychometric indices used to characterize items? The association between item accessibility and key psychometric indices was expected to be positive and to a moderate degree. Specifically, the relations between item accessibility as rated by the TAMI ARM and difficulty (*p*) was expected to result in a moderate correlation, while the correlation between ARM rating and item discrimination was expected to be low Both correlations, however, were expected to be of a magnitude that would be considered statistically significant The prediction was partially supported. Specifically, the correlations between ARM rating and difficulty for the two forms were very small and negative for Form A, and they were small to moderate for Form B. The correlations between ARM rating and item discrimination were moderate.

Question #3. How do students with different abilities perceive the accessibility of items, their cognitive demand, their teachers' coverage of the content, and their own predicted performance? Students' comprehension, cognitive ease, and perceived performance were expected to be significantly higher for students who received modified items compared to students who received original (unmodified) items. Specifically, student ratings of comprehension, cognitive ease, and perceived performance for the combined modified-accommodated (BA) and modified-nonaccommodated (BN) conditions were expected to be significantly higher than ratings for the combined original non-accommodated (AN) and original accommodated (AA) conditions based on results of unpaired *t* tests. In addition, comprehension, cognitive ease, and perceived performance are expected to be lower for students identified with disabilities than for their non-identified peers. Specifically, ratings of comprehension, cognitive

ease, and comprehension for students in the IEP group are expected to be be significantly lower than ratings for students with no IEPs based on results of unpaired *t* tests. Further, students who received accommodations, helpfulness of the accommodations was expected to be higher for students identified with disabilities than for their non-identified peers. The prediction was partially supported. Specifically, ratings of comprehension, cognitive ease, and perceived performance for students in the IEP group all were significantly lower than for students with no IEPs. Mean helpfulness of accommodations, however, was not significantly higher for students with IEPs, and desireability was low for both groups, based on the result that 93% of students did not avail themselves of the accommodations that were available to them.

Interpretation of Major Findings

I designed the current study to permit pairwise comparisons across four levels of access strategies: original items with no accommodations, original items with accommodations, modified items with no accommodations, and modified items with accommodations. No main effect of accommodations was detected however, likely due to the fact that only 7% of students requested and thus received the accommodations that were available to them. Notwithstanding, student survey data indicated students for whom accommodations were available reported they were helpful.

Because there was no detectable effect of accommodations on student scores, I collapsed the study design to examine the effect of modifications on the test scores of the total sample, and disaggregated by group. The main effect of modifications was low. Data, however, indicated a significant boost for students with IEPs compared to those without; the effect size for the IEP group was moderate and nearly double the size of that observed for students with no IEPs.

The reliability coefficients for Forms A and B were both high. Both forms were less reliable for students with IEPs than for students with no IEPs. The accessibility of the modified form was significantly higher than for the modified form, a finding that was consistent for all item elements with the exception of the item layout. For the original form, correlations between accessibility and item difficulty were very small and negative (range: -.04 to -.15); for the modified form, correlations were small to moderate (range: .03 to .30). Correlations between accessibility and item discrimination ranged from .07 to .42 for the total sample.

There are a number of plausible means of interpreting these mixed correlational results. First, the range of overall accessibility ratings using the ARM (Beddow et al., 2009) is 1-4; moreover, for Form A, all ratings were either 2 or 3; for Form B, there were no ratings below 3. Even if a relation between accessibility and item performance exists, the restricted range of the overall ratings for both forms likely reduced the magnitude of correlations in this study. Notwithstanding the fact that the range of the data would be increased if mean item element ratings were used instead of overall ratings, the technical manual for the instrument clearly indicates the overall rating is not simply a mathematical derivative of the element ratings. Further, the theoretical relation between accessibility and item performance is dependent on the degree of student mastery of the tested content; namely, if the an individual does not possess the intended construct of an item, no degree of modification to increase the accessibility of the item will increase the likelihood that he or she would respond correctly. Accessibility, therefore, is related to test performance only for students for whom access barriers may preclude them from demonstrating their knowledge of the tested content, and this relation is measurable only with the population who, barring accessibility barriers, would be able to demonstrate it. In the current

study, there likely is a subsample for each item to which these criteria apply, but the available data do not permit their identification.

Results of student survey data indicated student-reported comprehension was higher for the modified items compared to the original items. Students with IEPs reported lower comprehension than their general-education peers. Similarly, students reported they had to work less hard (i.e., lower cognitive load) on modified items compared to unmodified items. Again, students with IEPs reported they had to work harder on items compared to students with no IEPs. Students reported they had been taught the material to a greater degree for the modified items than the original items. Students with IEPs reported lower opportunity to learn the material compared to students with no IEPs. Students reported being more certain or sure they got the items correct for the modified versions compared to the unmodified versions. Students with IEPs reported they were significantly less sure about their performance on the items compared to students with no IEPs.

Comparison of Major Findings to Previous Research

Based on the extant literature on testing accommodations, the absence of an effect of accommodations on student test results is surprising, yet explainable. In their meta-analysis of accommodations research, Sireci and colleagues (2005) reported a main effect of the implementation of accommodations on student scores in all studies, and in studies that used an experimental design, an interaction (i.e., a differential boost for students with IEPs) was observed as well. Not only did I find no differential boost of accommodations by group notwithstanding the experimental design of the study, neither was there a main effect of

accommodations on student scores. The explanation of my findings seems to rest with the fact that very few students actually used the accommodations they were allowed.

Results of a study by Feldman, Kim, and Elliott (In press) on the provision of testing accommodations for adolescents on large-scale tests illuminate my explanation for the current study's findings. The data in that study indicated there was a main effect of accommodations on test scores in addition to a main effect of accommodations on students' self-perception (e.g., test self-efficacy and motivation.) Indeed, while the authors found no differential boost in test scores between students with and without IEPs, there was a significantly larger increase in self-efficacy and motivation for students in special education compared to their general education peers. The current study replicated Feldman et al.'s findings in terms of the main effects of accommodations on student-reported test self-efficacy; however, the actual provision of accommodations differed considerably (i.e., there was very little) and there was no actual effect of accommodations on test scores, and there was no interaction by student group.

It may well be that for any number of reasons, the junior high school student sample in the current study simply did not want accommodations, and therefore did not ask for them. Russell (In press) reported not only do many states find difficulty administering accommodations with integrity, but also that many students report accommodations are aversive (indeed, it is for this reason Russell has argued that accommodations should be delivered, to the extent possible, using computers as opposed to visible human proctors.) In the current study, it is possible junior high school students' aversion to using testing accommodations precluded the accommodations from producing their intended effect.

Using Cohen's (1988) suggested guidelines for interpreting the magnitude of effect sizes (i.e., .20 = small, .50 = medium, .80 = large), the modification effect for the current study may be

considered medium. However, Cohen indicated this guideline does not warrant strict adherence, to wit: “The terms 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation....In the face of this relativity, there is a certain risk inherent in offering conventional operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use only when no better basis for estimating the ES index is available” (p. 25).

In the current study, Bloom, Hill, Black, and Lipsey (2008) may provide a better frame of reference for interpreting the effects of the modifications on achievement test results. In their study of achievement performance trajectories, the mean effect size based on six nationally-normed tests indicated that student growth from grades 6 to 7 was .32 (range: .18 to .38, SD = .06). Based on these data, the effect size observed in the current study for students with IEPs is practically quite large – indeed, according to these findings, an effect size of .40 may represent an entire year’s growth. Thus, we may conclude that modifying a test with the singular purpose of removing access barriers while preserving the target construct may potentially reduce the observed achievement gap between students identified with, and not identified with, disabilities. Indeed, even if we take the difference between the two groups for the current study (i.e., .20) to represent the potential effect of applying accessibility theory to modify an existing test, the result would still reduce the gap in achievement to a practically significant degree.

The moderate correlation between item accessibility and item discrimination supports Beddow, Kurz, and Frey’s (In press) argument that the accessibility of a test influences the

degree to which the results will distinguish between those who know the tested content and those who do not. Likewise, the correlation with item difficulty supports the argument that access barriers may preclude some test-takers from demonstrating fully what they know and can do on a test.

Limitations

There are several aspects of the study design that limit the generalizations and implications that can be derived from the results. First, the study used mathematics items only. Elliott et al. (2009) found larger effect sizes for reading modifications, suggesting common modification strategies such as reducing reading load have a greater impact on reading items. Second, the participants were from one grade, so results cannot necessarily be generalized across the grade span. Fourth, notwithstanding the fact that the inclusion of English Language Learners (ELLs) and students with limited English proficiency (LEPs) into the study was likely to introduce variability into test results on account of their language abilities and educational histories, these important populations were not specifically recruited for the current study, nor were they excluded from analyses. In theory, these students are likely to benefit from the types of accessibility strategies used in this study, but these students were not isolated to permit the examination of accommodations and modifications on their test results. It should also be noted the study design did not permit within-group analyses, which likely resulted in greater unexplained variability in results.

Additionally, the current study was limited in that the test forms consisted solely of multiple-choice items. There are implications based on accessibility theory for modifying other item types to increase their accessibility for more test-takers (e.g., constructed response items),

and this is important insofar as a large number of tests use other types of items. Further, the study used test forms that created for the purpose of this study, so results could not be applied to a proficiency standard. Specifically, there are not cut scores to compare against student performance, so the potential impact of results on proficiency rates could not be ascertained.

Further, the modified test form was not field-tested prior to their use in this study. All of the items for Form A were vetted through Discovery Education Assessment's historical item statistics, but the investigator for the study designed the items for Form B with the aim of integrating all of the suggestions of the expert accessibility rating team. Generally, reliability of test scores is considered a necessary condition for the validity of subsequent inferences, and the reliability of Form B was considerably lower than that of Form A, an indication that there were some problematic items. The web-based test-delivery system used by DEA posed challenges in designing modified versions of some of items, resulting in some items that required test-takers to scroll to see one or more of the answer choices. This scrolling introduced an element of cognitive demand that was not present in the original test items. The magnitude of the effect of representational holding on student test scores was not measurable within the framework of the study. In a typical test development cycle, these items would have been discovered during field-testing and modified or replaced.

The execution of the study was also limited in several ways. First, the desired sample size was not obtained, specifically with regard to students with IEPs. This precluded the application of exclusion criteria to this group (e.g., ensuring student participants all were identified with SLD in reading or mathematics and had scored below proficient on the previous two years' assessments in mathematics.) Second, the lack of student response to the availability of testing accommodations was unanticipated; notwithstanding the effort to administer accommodations as

authentically as possible, the decision to deliver testing accommodations to every participant in the accommodated conditions may have been preferable.

Student responses about the items were obtained following the test itself rather than during the actual test-taker response process. As a result, the validity of the inferences made from perceived test-taker access, coverage, and difficulty data may be threatened because of test-taker fatigue and/or recall failure. Additionally, this feedback was solicited in a group setting, where peer influence may have threatened the validity of the results. This limitation was addressed insofar as student responses were kept confidential during the group sessions, but it was not always possible to prohibit students from commenting aloud about the test items. Additionally, time constraints and concerns about participant fatigue necessitated collecting survey data about a subsample consisting of 18 of the 68 items (selected at random), which limited the generalizability of student and teacher responses to the larger item sample. Finally, only 4 of the 17 participating teachers completed the online teacher survey. It behooves the testing community to incorporate teacher and student voices into the research on accessibility and test development, because notwithstanding the end users of the tests are students, teachers often are included in the development process.

Implications of Current Findings

The current study supports the argument that accessibility is directly related to the validity of inferences that can be made from achievement test scores. Item accessibility is a construct that has remained largely unmeasured, and the results of this study indicate there are several relations between accessibility and validity-related variables including item difficulty, item discrimination, readability, student test self-efficacy, and ultimately student proficiency.

Further, it appears students who are not identified with disabilities also are affected by these relations; indeed, the data suggest there are students in the general education population for whom some test items do not permit unfettered access to the target construct.

Several basic features of the items were changed as well, including a 26% reduction in word count by item and a 33% reduction in words for the overall form. Additionally, the readability scores for the items were lower for Form B. Moreover, the changes in font sizes and increase in white space increased the legibility of the test as a whole. These features, while not considered critical by test developers, may hold greater promise for change than many other aspects of current tests. Indeed, I noted that the correlation between the difference in word counts between the two items and the change in item difficulty was -.30, indicating that student performance on the math test was related to the number of words the test required the students to read.

Overall, these data suggest that while we have much to learn about how to identify and address accessibility concerns in individual items, it would behoove test developers to apply the accessibility review process to current achievement tests with the aim of removing access barriers that may reduce their capacity to measure what students know and can do.

Directions for Future Research

There are several important aspects of accessibility theory that remain for examination by future researchers. First, while the current study provided some opportunity for students to report on their perceived ability to access the items on the field test, future researchers should consider the use of computer-based test delivery systems to solicit student feedback about the items during the test itself, as well as to record the amount of time spent by the test-taker on each item.

This would permit a closer examination of the relations among accessibility and cognitive load proxies including time spent, perceived difficulty and complexity, and actual test-taker performance. Second, notwithstanding the variety of issues with administering accommodations to students who are not eligible for them, or for whom they are not desired (e.g., distracting them from performing optimally, causing frustration, etc.) the study may have benefited from delivering a specific package of accommodations to each student regardless of need or desire. Third, the current study should be replicated with a more robust sample including a representative students at multiple grade-levels identified with, and not identified with, disabilities, as well as with ELL and LEP student populations. This is not easy work without substantial funding, but important if we are to continue to move forward to improve test accessibility for all students.

Conclusion

Test accessibility is a relatively new but potentially critical psychometric construct of tests and a test event. For a test to yield results from which valid inferences can be made, the test must be accessible for the test-takers to whom the inferences will be applied. The current study advanced the knowledge of how test accessibility may function as part of an overall validity argument, especially for students with disabilities and perhaps for English language learners as well. Test developers should consider integrating item accessibility reviews into the test development process, to ensure all tests permit the entirety of the test-taker population to show what they know and can do. More accessible tests will make the test event more meaningful for students and others who use test scores to make important decisions about the students.

FOOTNOTES

[1]Consortium for Modified Alternate Assessment Development and Implementation (CMAADI). CMAADI is directed by Stephen N. Elliott, Michael C. Rodriguez, Andrew T. Roach, and Ryan J. Kettler. The project is funded by the U.S. Department of Education, Office of Special Education and Rehabilitative Services and involves the Arizona and Indiana Departments of Education.

[2]Although the effect size for modifications was smaller in early work in this area (i.e., the CAAVES project; Elliott et al., 2009), I hypothesized the extant results likely reflected error due to order and form effects because of the within-groups design used for the study. Additionally, the modifications in that study were conducted by item writers with minimal training, resulting in less consistency and less robust modifications. Additionally, numerous changes were made to the original items, which likely attenuated the observed differences. Based on this rationale, a main effect size of .50 was a reasonable expectation for each of the three experimental conditions against the control group for the current study.

Appendix A

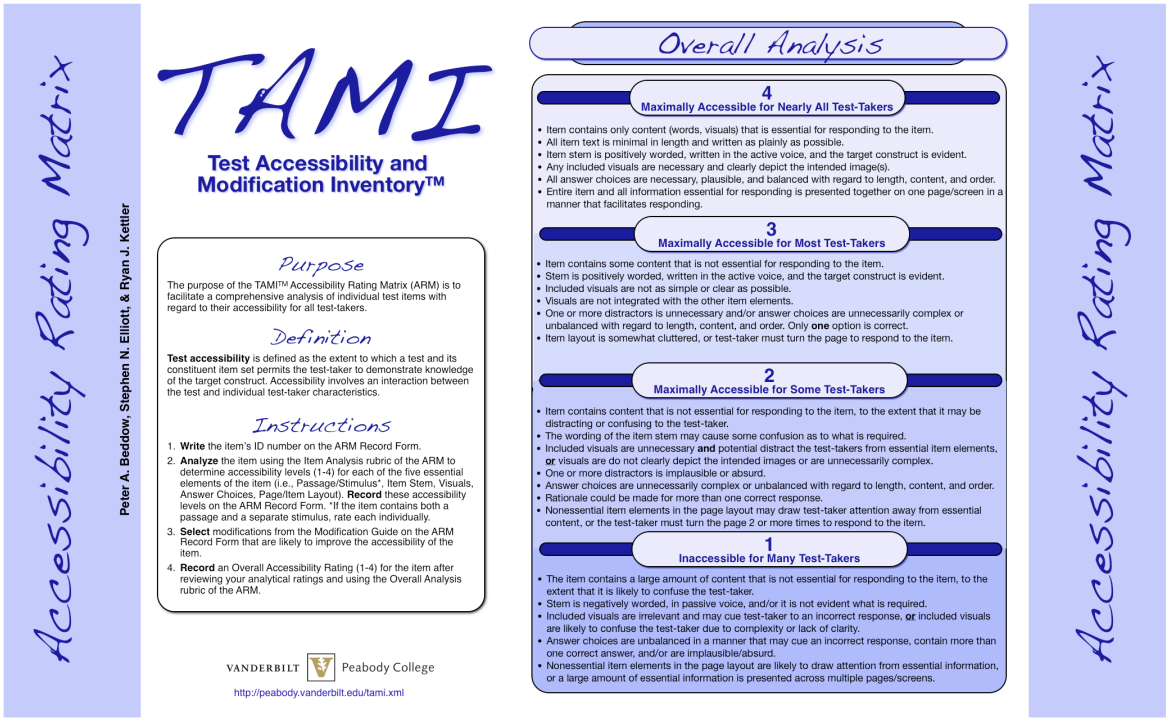


Figure A1. TAMI Accessibility Rating Matrix, page 1.

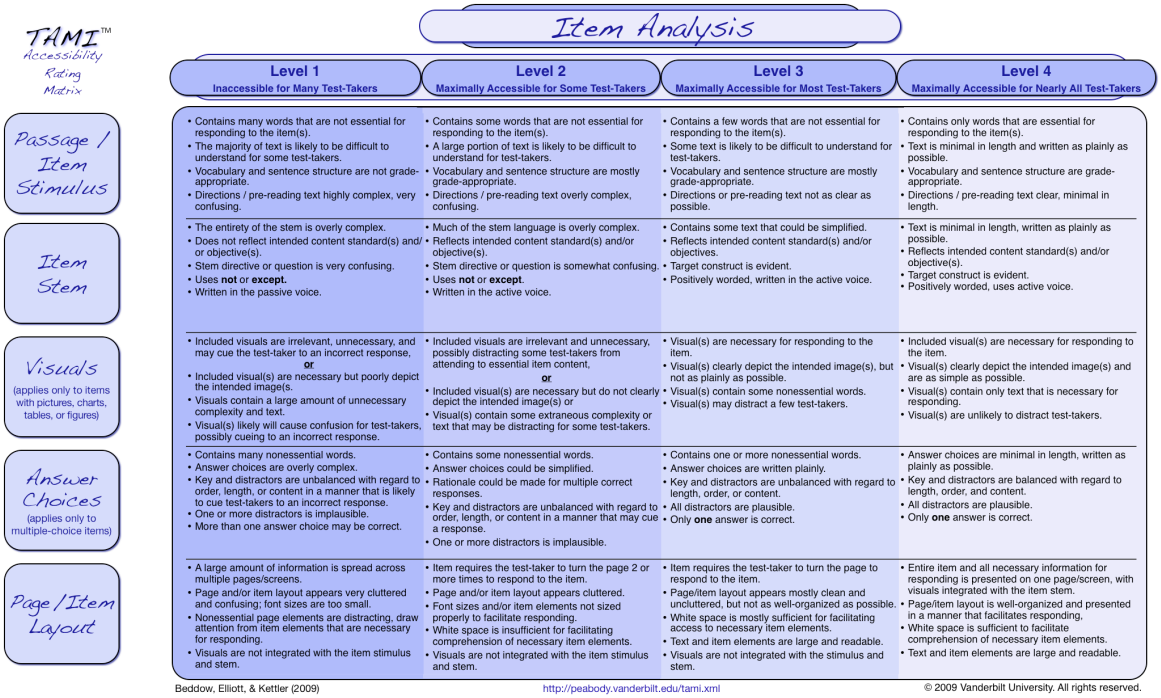


Figure A2. TAMI Accessibility Rating Matrix, page 2

.

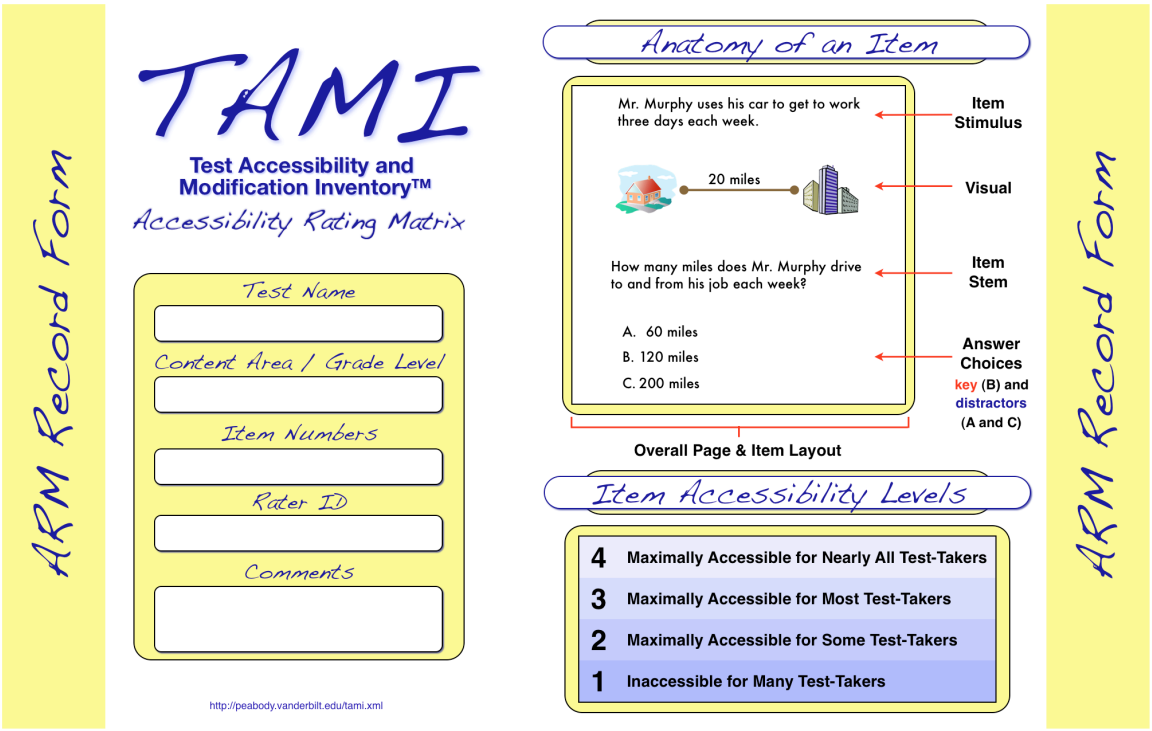


Figure A3. TAMI Accessibility Rating Matrix, page 3.

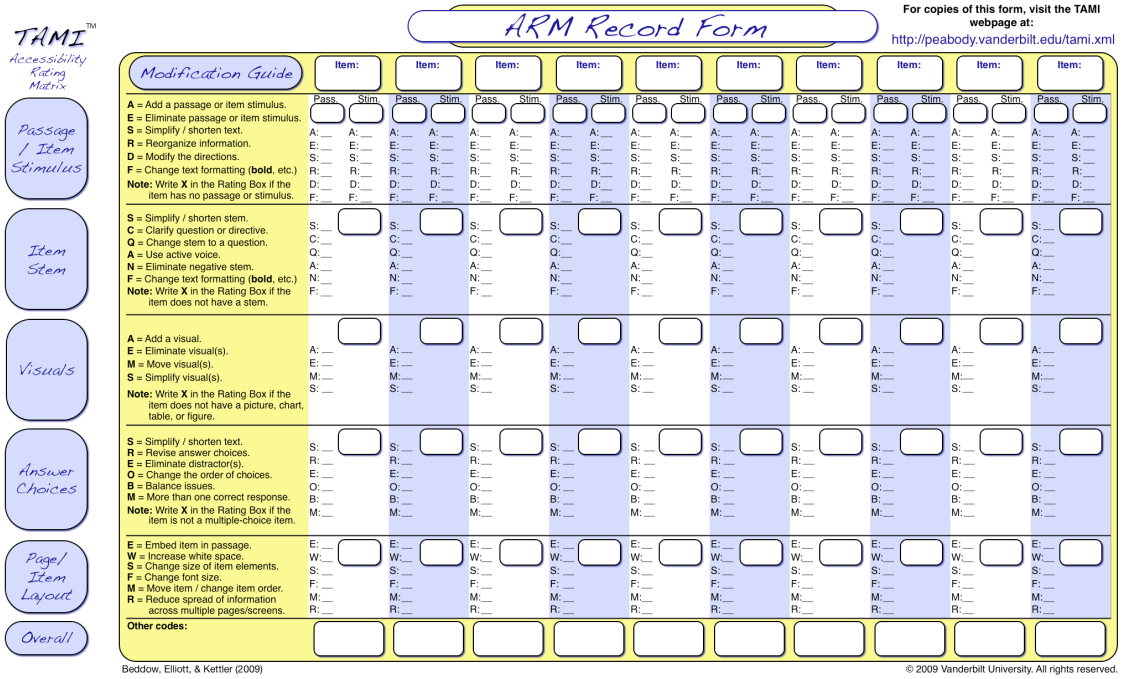
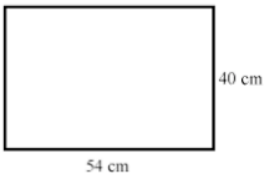


Figure A4. TAMI Accessibility Rating Matrix, page 4.

Appendix B

1. **DIRECTIONS:** There are 34 questions on this test. Please answer every question.

A student's desktop measures 54 cm long and 40 cm wide.



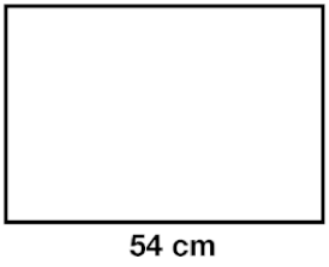
What is the perimeter of the desktop?

A. 2,160 centimeters
B. 216 centimeters
C. 188 centimeters
D. 94 centimeters

1.

Directions: There are 34 questions on this math test. Please do your best and answer every question.

1 A desktop measures 54 cm long and 40 cm wide.



What is the **perimeter?**

A. 94 cm

B. 188 cm

C. 2160 cm

Figure B1. Item #1 in original and modified forms.

Table B1.
Item Statistics for Item #1

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.58**	**0.60**	+.02
	No IEP	0.63	0.66	+.03
	IEP	0.44	0.43	-0.01
Discrimination (*D*)	**Total**	**0.49**	**0.40**	**-0.09**
	No IEP	0.46	0.48	+0.02
	IEP	0.54	0	-0.54
Accessibility		**3**	**3**	**0**
Readability		4.5	2.2	-1.3
Word Count		18	14	-4
Student Variables				
Comprehension [1]	**Total**	**4.9 (1.4)**	**5.0 (1.3)**	**+0.1**
	No IEP	5.0 (1.2)	5.4 (1.1)	+0.4
	IEP	4.5 (1.9)	3.9 (1.4)	-0.6
Cognitive Demand [1]	**Total**	**2.6 (1.3)**	**2.4 (1.4)**	**-0.2**
	No IEP	2.4 (1.1)	2.2 (1.3)	-0.2
	IEP	3.0 (1.7)	2.8 (1.5)	-0.2
OTL [1]	**Total**	**4.2 (1.6)**	**4.3 (1.4)**	**+0.1**
	No IEP	4.4 (1.5)	4.6 (1.2)	+0.2
	IEP	3.7 (1.9)	3.3 (1.4)	-0.4
Confidence [1]	**Total**	**4.3 (1.6)**	**4.7 (1.7)**	**+0.4**
	No IEP	4.4 (1.5)	4.9 (1.5)	+0.5
	IEP	4.2 (1.9)	3.8 (1.9)	-0.4
Help [1]	**Total**	**2.7 (1.9)**	**3.5 (1.9)**	**+0.8**
	No IEP	2.6 (1.9)	3.3 (2.1)	+0.5
	IEP	2.8 (2.0)	3.8 (1.2)	+1.0

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

2. On family night at the carnival, there was a sign showing a special on tickets for rides.

Number of Rides Purchased (x)	Number of Free Rides (y)
2	1
4	2
6	3
8	4
10	5

Which equation can be used to find the number of free rides?

A. $y = \frac{1}{2}x$

B. $y = 2x$

C. $y = \frac{1}{3}x$

D. $y = 3x$

2.

2 Sara went to the Fair.

This sign showed the prices for ride tickets.

Tickets purchased (x)	Free tickets (y)
2	1
4	2
6	3

How does y relate to x?

A. $y = \frac{1}{2}x$

B. $y = 2x$

C. $y = 3x$

Figure B2. Item #2 in original and modified forms.

Table B2.
Item Statistics for Item #2

Item Variables		Original	Modified	Δ
Difficulty (p)	**Total**	**0.45**	**0.49**	**+.04**
	No IEP	0.45	0.51	+.06
	IEP	0.46	0.43	-0.03
Discrimination (D)	**Total**	**0.25**	**0.29**	**+0.04**
	No IEP	0.35	0.36	+0.01
	IEP	0.15	0.33	+0.18
Accessibility		**3**	**3**	**0**
Readability		7.3	2.1	-5.2
Word Count		29	19	-10

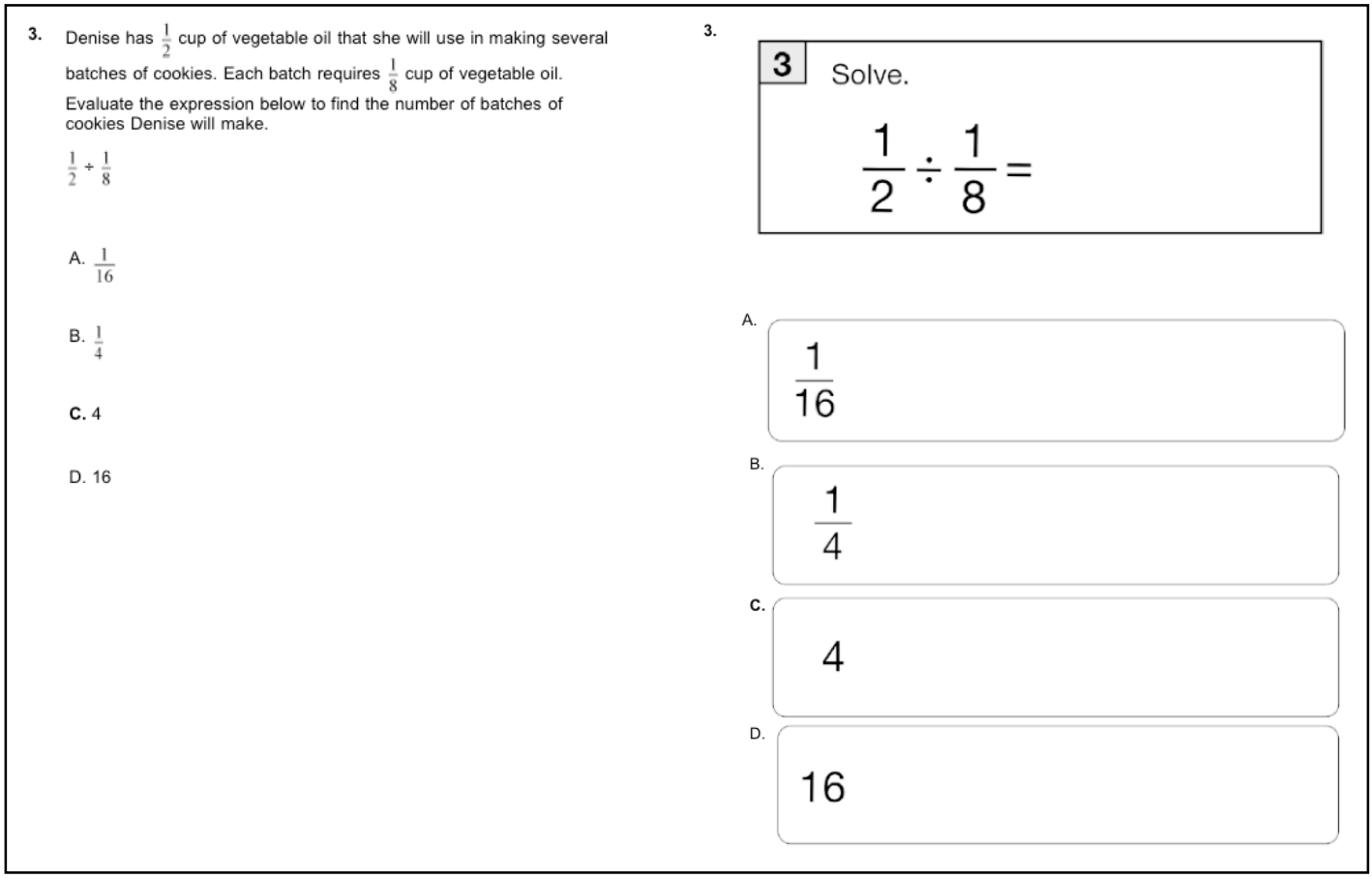


Figure B3. Item #3 in original and modified forms.

Table B3.
Item Statistics for Item #3

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.51**	**0.42**	**-0.09**
	No IEP	0.58	0.49	-0.09
	IEP	0.33	0.18	-0.17
Discrimination (*D*)	**Total**	**0.63**	**0.58**	**-0.04**
	No IEP	0.51	0.52	+0.02
	IEP	0.23	0.50	+0.27
Accessibility		**2**	**3**	**+1**
Readability		10.0	5.3	-4.7
Word Count		36	1	-35
Student Variables				
Comprehension [1]	**Total**	**4.8 (1.5)**	**5.0 (1.6)**	**+0.2**
	No IEP	4.9 (1.4)	5.1 (1.5)	+0.2
	IEP	4.3 (1.6)	4.9 (1.7)	+0.6
Cognitive Demand [1]	**Total**	**2.5 (1.4)**	**2.5 (1.6)**	**0.0**
	No IEP	2.4 (1.3)	2.4 (1.6)	0.0
	IEP	2.8 (1.7)	2.6 (1.4)	-0.2
OTL [1]	**Total**	**4.5 (1.4)**	**4.6 (1.3)**	**+0.1**
	No IEP	4.7 (1.2)	4.6 (1.3)	-0.1
	IEP	3.8 (1.6)	4.3 (1.4)	+0.5
Confidence [1]	**Total**	**4.4 (1.7)**	**4.6 (1.7)**	**+0.2**
	No IEP	4.6 (1.6)	4.7 (1.8)	+0.1
	IEP	3.8 (1.9)	4.3 (1.6)	+0.5
Help [1]	**Total**	**3.0 (2.1)**	**3.5 (2.0)**	**+0.5**
	No IEP	2.8 (2.0)	3.4 (2.1)	+0.6
	IEP	3.4 (2.5)	3.8 (1.9)	+0.4

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

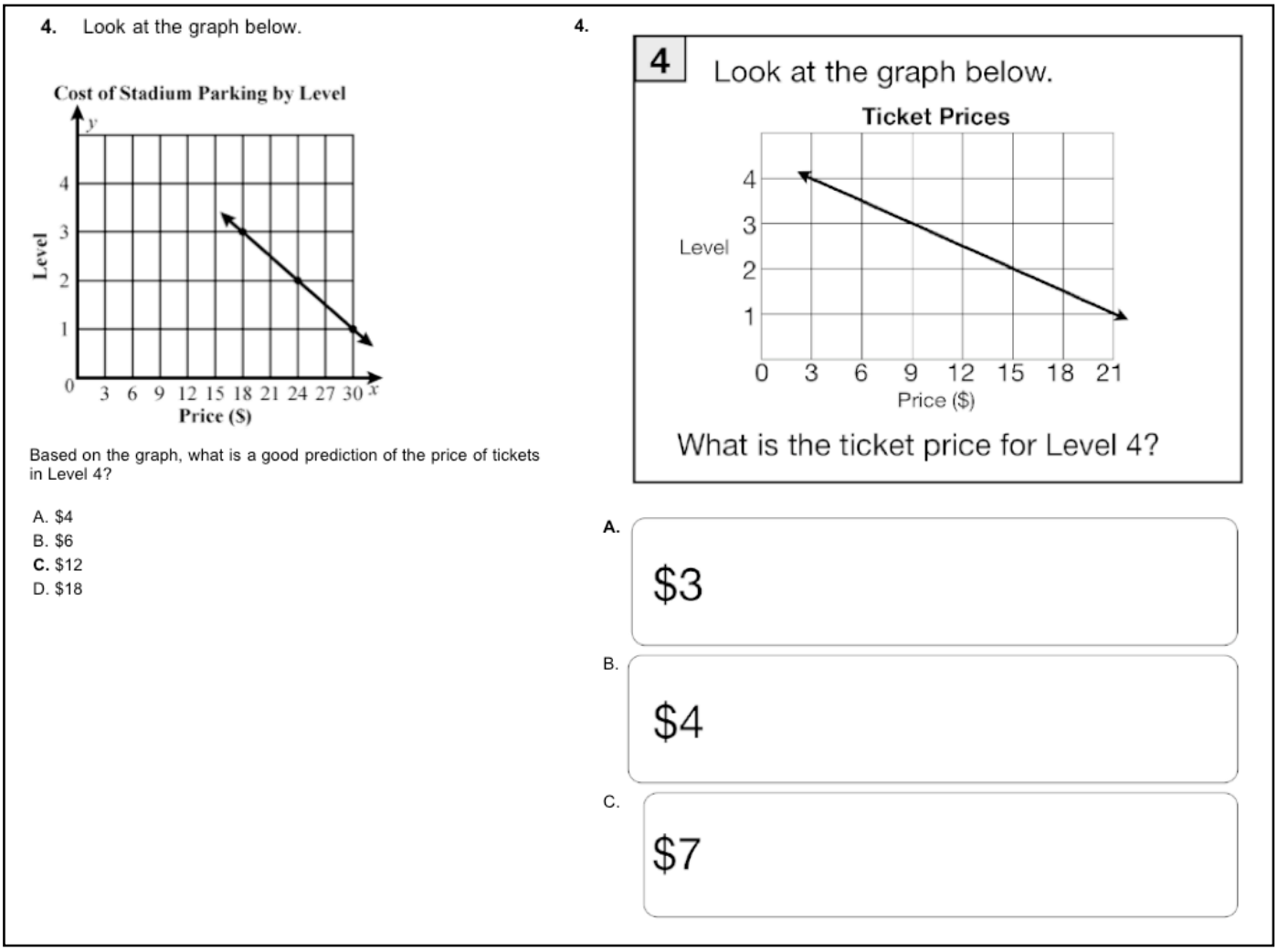


Figure B4. Item #4 in original and modified forms.

Table B4.
Item Statistics for Item #4

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.76**	**0.85**	**+0.09**
	No IEP	0.81	0.89	+0.08
	IEP	0.62	0.75	+0.13
Discrimination (*D*)	**Total**	**0.47**	**0.18**	**-0.29**
	No IEP	0.30	0.12	-0.18
	IEP	0.69	0.42	+0.28
Accessibility		**3**	**4**	**+1**
Readability		4.2	2.0	-2.2
Word Count		22	17	-5

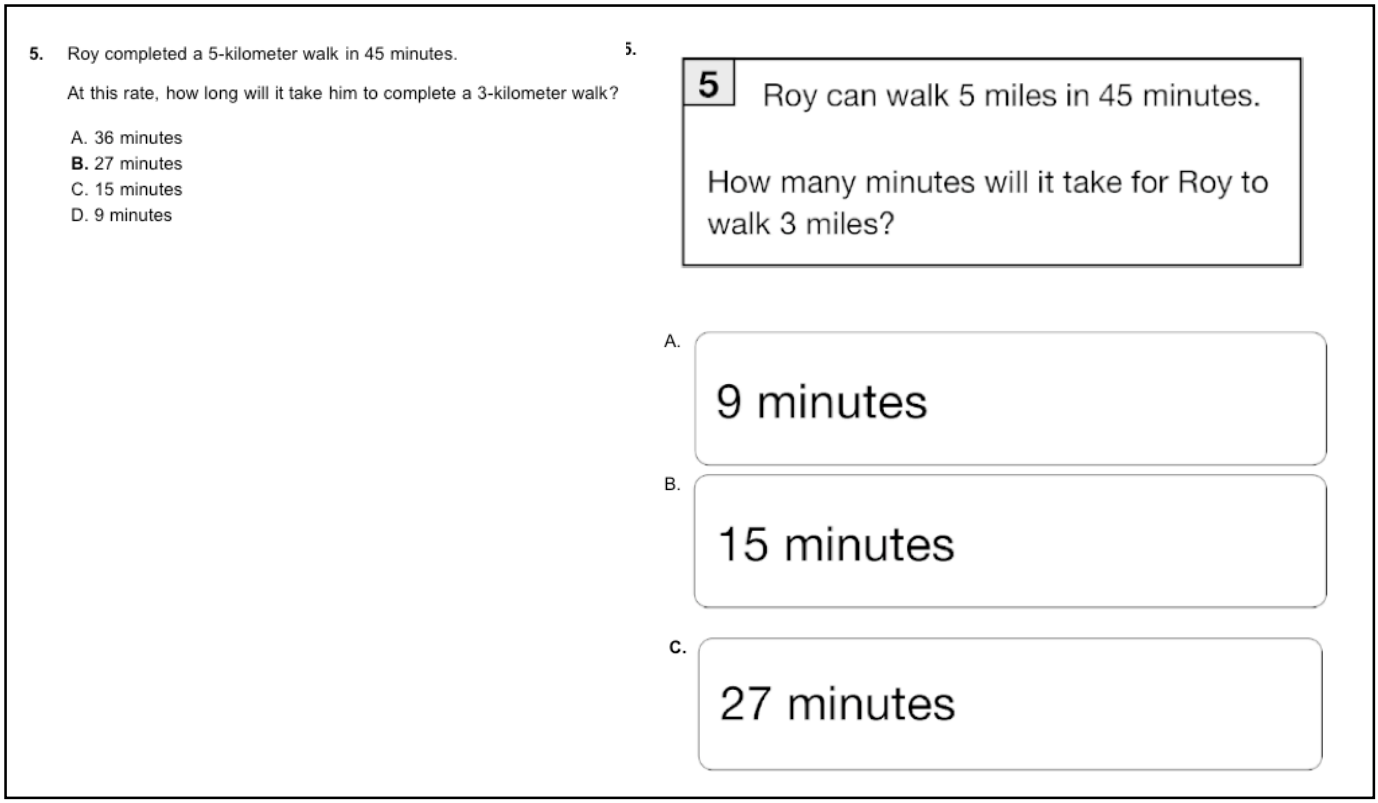


Figure B5. Item #5 in original and modified forms.

Table B5.
Item Statistics for Item #5

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.56**	**0.57**	**+0.01**
	No IEP	0.60	0.65	+0.05
	IEP	0.44	0.31	-0.15
Discrimination (*D*)	**Total**	**0.55**	**0.62**	**0.07**
	No IEP	0.56	0.55	-0.01
	IEP	0.38	0.08	-0.30
Accessibility		**3**	**4**	**+1**
Readability		6.2	2.0	-4.2
Word Count		24	20	-4
Student Variables				
Comprehension [1]	**Total**	**4.3 (1.8)**	**4.6 (1.5)**	**+0.3**
	No IEP	4.5 (1.6)	4.8 (1.5)	+0.3
	IEP	3.6 (2.0)	4.3 (1.6)	+0.7
Cognitive Demand [1]	**Total**	**2.8 (1.6)**	**3.1 (1.5)**	**+0.3**
	No IEP	2.8 (1.6)	3.1 (1.6)	+0.3
	IEP	2.8 (1.7)	3.3 (1.5)	+0.5
OTL [1]	**Total**	**3.8 (1.7)**	**3.6 (1.4)**	**-0.2**
	No IEP	3.9 (1.6)	3.6 (1.5)	-0.3
	IEP	3.4 (1.8)	3.5 (1.0)	+0.1
Confidence [1]	**Total**	**4.0 (1.7)**	**4.2 (1.8)**	**+0.2**
	No IEP	4.1 (1.7)	4.3 (1.7)	+0.2
	IEP	3.9 (1.9)	3.8 (2.0)	-0.1
Help [1]	**Total**	**2.9 (2.0)**	**3.5 (1.9)**	**+0.6**
	No IEP	2.8 (1.9)	3.4 (2.0)	+0.6
	IEP	3.2 (2.6)	3.6 (1.7)	+0.4

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

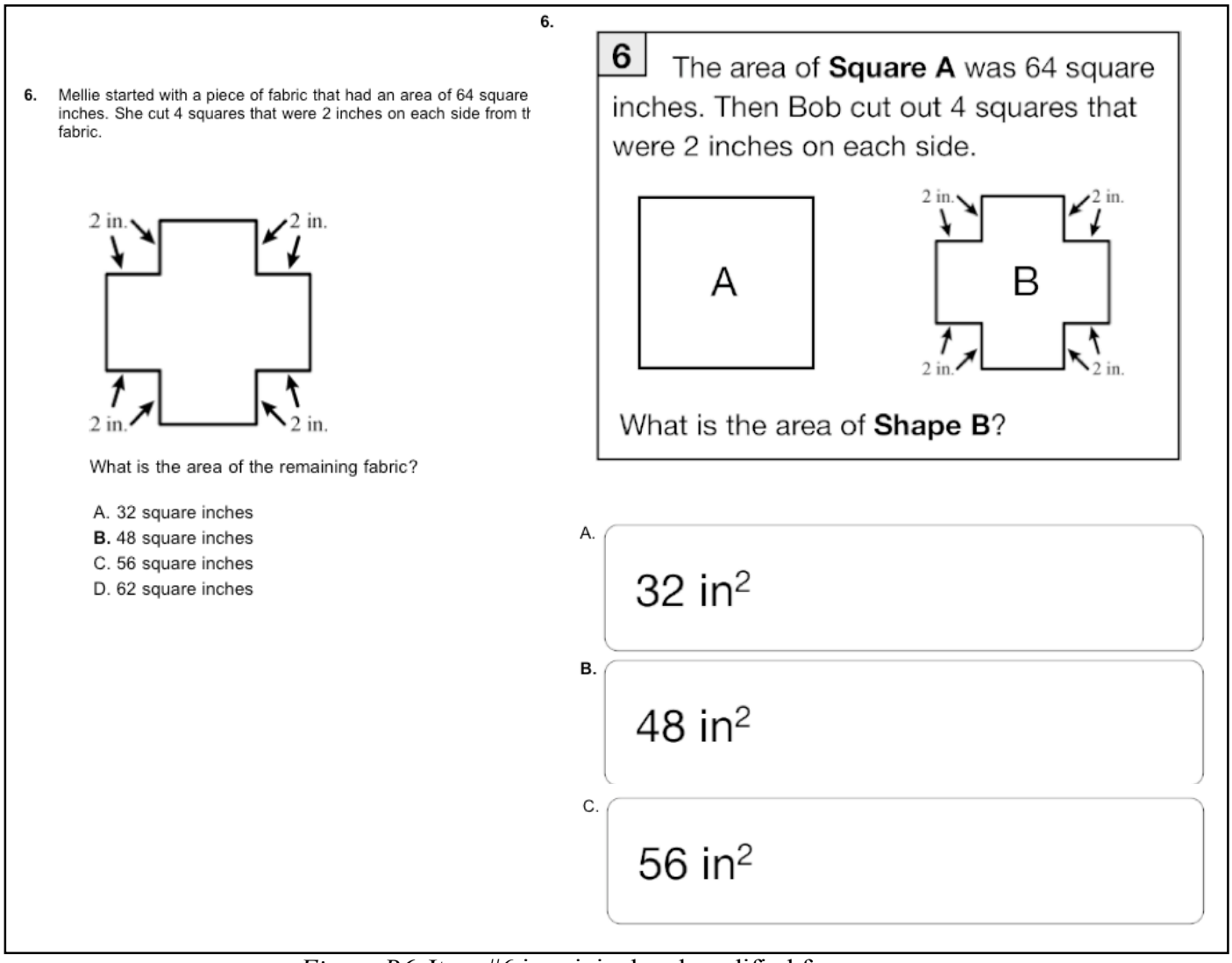


Figure B6. Item #6 in original and modified forms.

Table B6.
Item Statistics for Item #6

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.32**	**0.42**	**+0.10**
	No IEP	0.33	0.41	+0.09
	IEP	0.29	0.47	+0.18
Discrimination (*D*)	**Total**	**0.18**	**0.18**	**0.00**
	No IEP	0.33	0.26	-0.07
	IEP	-0.38	0.08	+0.47
Accessibility		**2**	**3**	**+1**
Readability		4.6	1.8	-2.8
Word Count		37	29	-8

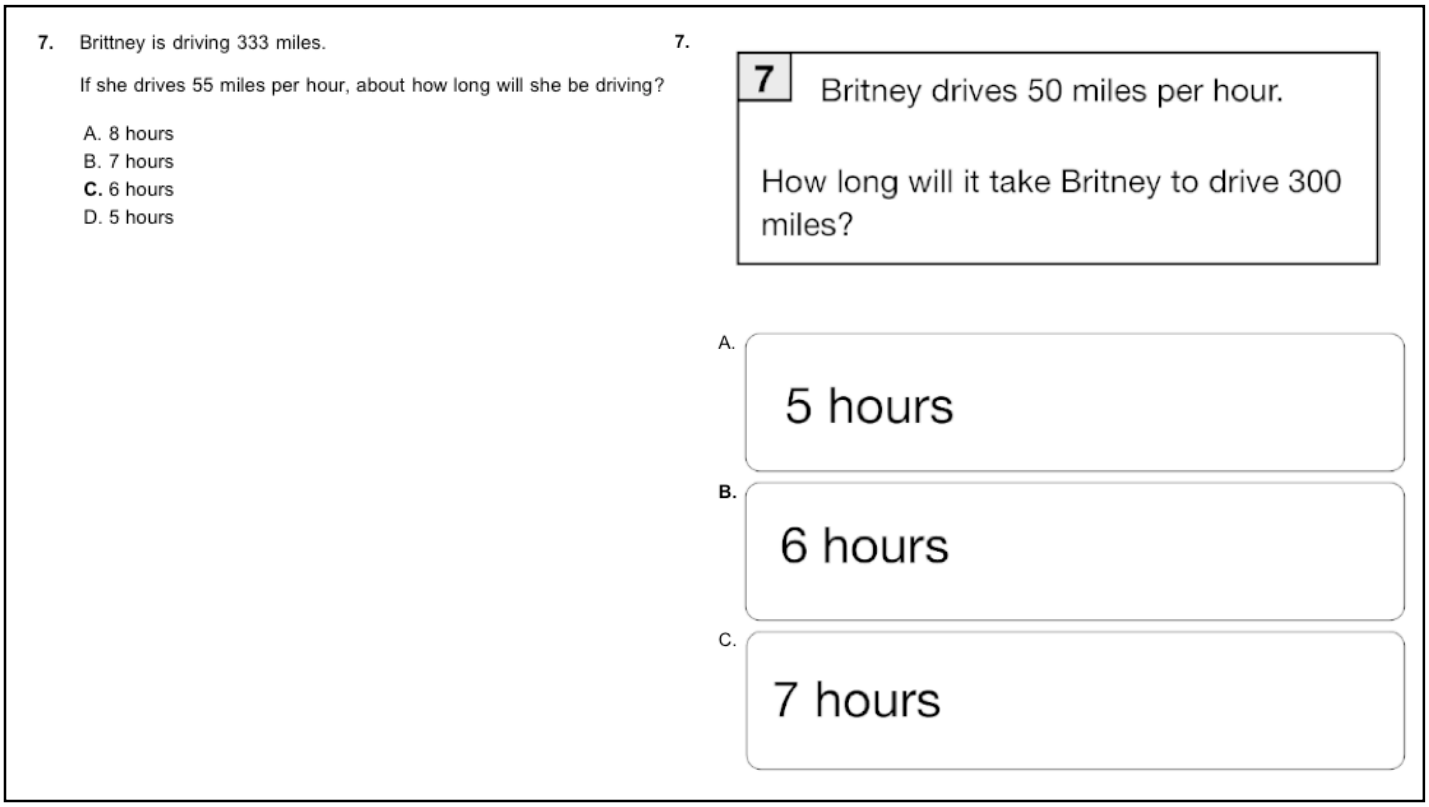


Figure B7. Item #7 in original and modified forms.

Table B7.
Item Statistics for Item #7

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.64**	**0.88**	**+0.24**
	No IEP	0.72	0.91	+0.19
	IEP	0.38	0.78	+0.40
Discrimination (*D*)	**Total**	**0.57**	**0.31**	**-0.26**
	No IEP	0.20	0.21	+0.01
	IEP	0.38	0.33	-0.05
Accessibility		**3**	**4**	**+1**
Readability		3.2	2.6	-0.6
Word Count		19	16	-3

8. Henry has a dog-sitting business. Each day, he charges \$10 for his time, plus \$5 per dog. This is shown by the equation below, where t is the total cost, in dollars, per day and n is the number of dogs.

$t = 10 + (5 \times n)$

Which chart shows Henry's prices?

A.

Number of Dogs (n)	Total Cost per Day (t)
1	\$5
2	\$10
3	\$15
4	\$20
5	\$25

B.

Number of Dogs (n)	Total Cost per Day (t)
1	\$10
2	\$20
3	\$30
4	\$40
5	\$50

C.

Number of Dogs (n)	Total Cost per Day (t)
1	\$15
2	\$30
3	\$45
4	\$60
5	\$75

D.

Number of Dogs (n)	Total Cost per Day (t)
1	\$15
2	\$20
3	\$25
4	\$30
5	\$35

8.

8 Henry walks dogs. Each day, he charges \$10 for his time, plus \$5 per dog. This is shown by the equation below.

$t = 10 + (5 \times n)$

t = total cost
n = number of dogs

Which chart shows Henry's prices?

A.

Number of Dogs (n)	Total Cost (t)
1	\$5
2	\$10
3	\$15

B.

Number of Dogs (n)	Total Cost (t)
1	\$15
2	\$20
3	\$25

C.

Number of Dogs (n)	Total Cost (t)
1	\$15
2	\$30
3	\$45

Figure B8. Item #8 in original and modified forms.

Table B8.
Item Statistics for Item #8

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.25**	**0.49**	**+0.24**
	No IEP	0.30	0.51	+0.21
	IEP	0.12	0.41	+0.29
Discrimination (*D*)	**Total**	**0.61**	**0.33**	**-0.28**
	No IEP	0.48	0.29	-0.20
	IEP	0.38	0.08	-0.30
Accessibility		**2**	**3**	**+1**
Readability		4.5	3.3	-1.2
Word Count		47	27	-20
Student Variables				
Comprehension [1]	**Total**	**4.1 (1.6)**	**4.8 (1.2)**	**+0.7**
	No IEP	4.1 (1.7)	5.1 (1.0)	+1.0
	IEP	3.8 (1.4)	3.9 (1.3)	+0.1
Cognitive Demand [1]	**Total**	**3.2 (1.4)**	**2.7 (1.5)**	**-0.5**
	No IEP	3.1 (1.4)	2.5 (1.5)	-0.6
	IEP	3.3 (1.3)	3.5 (1.5)	+0.2
OTL [1]	**Total**	**3.5 (1.5)**	**4.1 (1.4)**	**+0.6**
	No IEP	3.6 (1.5)	4.2 (1.3)	+0.6
	IEP	3.3 (1.6)	3.6 (1.7)	+0.3
Confidence [1]	**Total**	**3.9 (1.6)**	**4.5 (1.6)**	**+0.6**
	No IEP	4.1 (1.7)	4.9 (1.2)	+0.8
	IEP	3.7 (1.3)	3.0 (1.8)	-0.7
Help [1]	**Total**	**2.1 (1.6)**	**3.6 (2.3)**	**+1.5**
	No IEP	1.8 (1.3)	3.3 (2.3)	+1.5
	IEP	2.8 (2.1)	5.5 (0.7)	+2.7

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

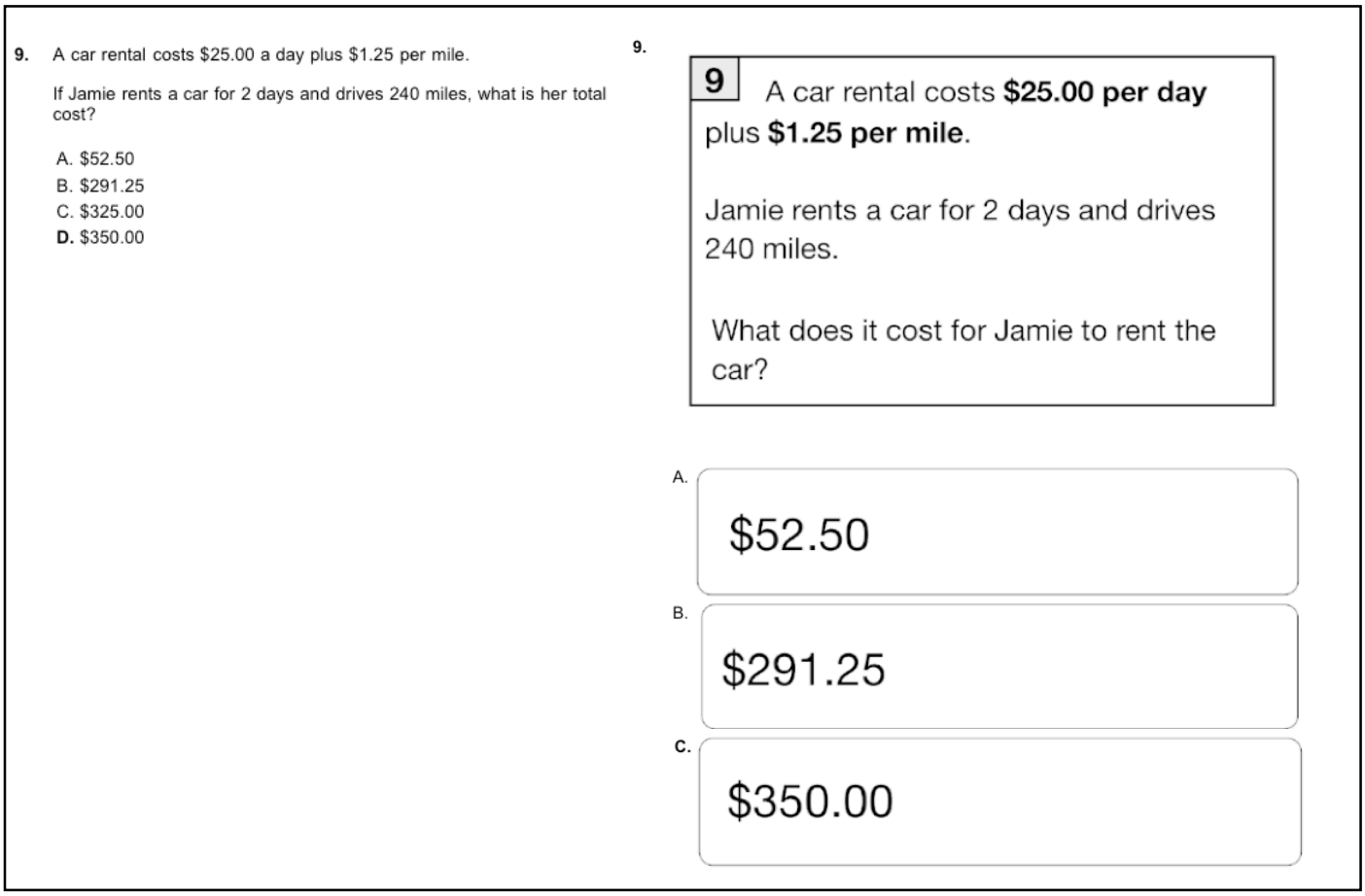


Figure B9. Item #9 in original and modified forms.

Table B9.
Item Statistics for Item #9

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.43**	**0.62**	**+0.19**
	No IEP	0.49	0.71	+0.22
	IEP	0.25	0.33	+0.08
Discrimination (*D*)	**Total**	**0.59**	**0.47**	**-0.12**
	No IEP	0.38	0.33	-0.05
	IEP	0.15	0.67	+0.51
Accessibility		**3**	**4**	**+1**
Readability		0.3	0.4	+0.1
Word Count		30	34	+4

10. Benjamin recorded the depth of the water in the bay throughout the day. His first measurement showed that the water was 12 feet deep. The next three measurements showed that the water had risen 1 foot, dropped 2 feet, and risen 1 foot again.

Which expression can be used to find the height of the water after Benjamin's last measurement?

A. 12 + 1 + 2 + 1
B. 12 + 1 + (-2) + 1
C. 12 + (-1) + 2 + 1
D. 12 + (-1) + (-2) + (-1)

10.

10 Ben measured the water level in the lake three times:

1. The water level was 12 feet.
2. The water level **rose** by 1 foot.
3. The water level **dropped** by 2 feet.

Which expression shows Ben's measurements?

A. 12 + 1 + 2

B. 12 + 1 + (-2)

C. 12 + (-1) + (-2)

D. 12 + (-1) + 2

Figure B10. Item #10 in original and modified forms.

Table B10.
Item Statistics for Item #10

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.80**	**0.90**	**+0.10**
	No IEP	0.86	.093	+0.07
	IEP	0.63	0.80	+0.17
Discrimination (*D*)	**Total**	**0.43**	**0.24**	**-0.20**
	No IEP	0.21	0.17	-0.04
	IEP	0.54	0.33	-0.21
Accessibility		**2**	**3**	**+1**
Readability		8.0	3.5	-4.5
Word Count		60	38	-22
Student Variables				
Comprehension [1]	**Total**	**4.5 (1.6)**	**4.9 (1.3)**	**+0.4**
	No IEP	4.6 (1.5)	5.2 (1.0)	+0.6
	IEP	4.2 (1.9)	4.3 (1.8)	+0.1
Cognitive Demand [1]	**Total**	**2.9 (1.5)**	**2.5 (1.4)**	**-0.4**
	No IEP	2.8 (1.4)	2.2 (1.3)	-0.6
	IEP	3.3 (1.9)	3.3 (1.4)	0.0
OTL [1]	**Total**	**3.9 (1.6)**	**4.0 (1.5)**	**+0.1**
	No IEP	3.9 (1.6)	4.1 (1.5)	+0.2
	IEP	3.7 (1.8)	3.8 (1.6)	+0.1
Confidence [1]	**Total**	**4.0 (1.5)**	**4.9 (1.3)**	**+0.9**
	No IEP	4.2 (1.6)	5.1 (1.1)	+0.9
	IEP	3.7 (1.4)	4.3 (1.7)	+0.6
Help [1]	**Total**	**1.9 (1.4)**	**3.7 (2.3)**	**+1.8**
	No IEP	1.6 (0.9)	3.5 (2.4)	+1.9
	IEP	2.8 (2.1)	5.0 (0.0)	+2.2

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

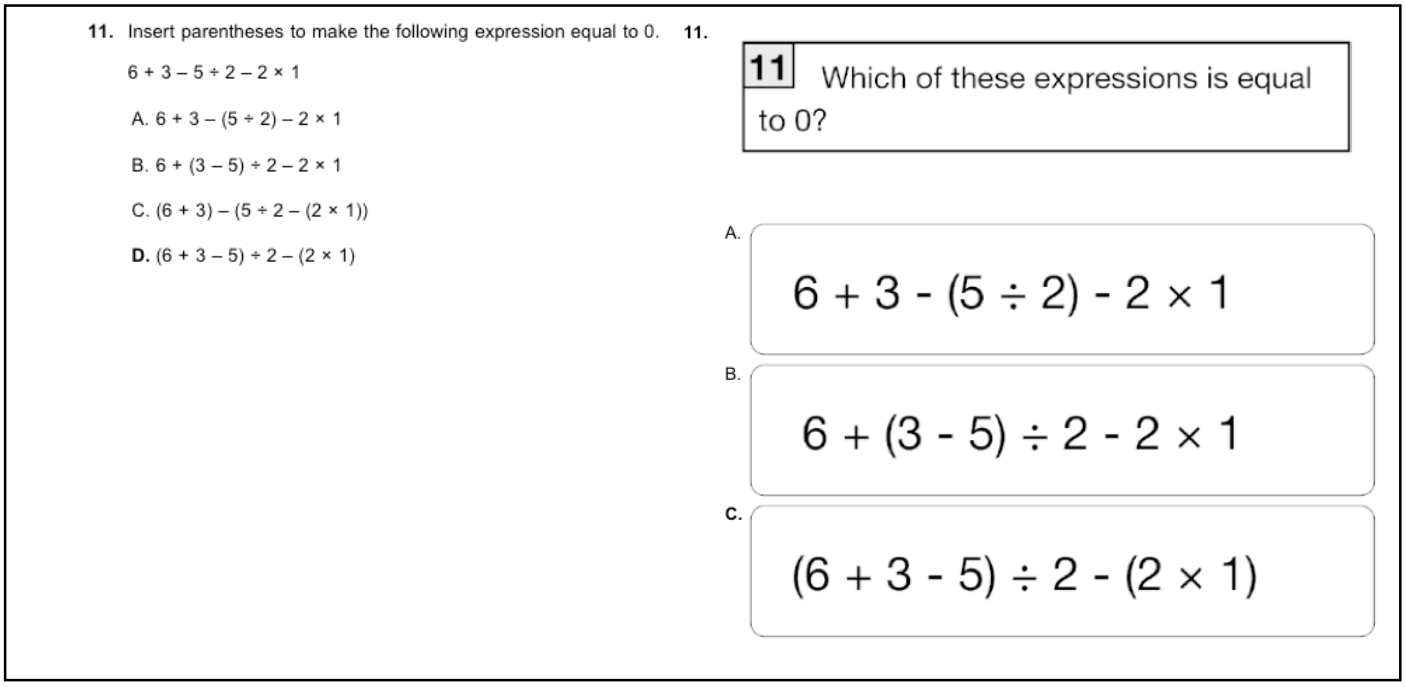


Figure B11. Item #11 in original and modified forms.

Table B11.
Item Statistics for Item #11

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.55**	**0.75**	**+0.20**
	No IEP	0.63	0.79	+0.16
	IEP	0.31	0.61	+0.30
Discrimination (*D*)	**Total**	**0.44**	**0.31**	**-0.13**
	No IEP	0.27	0.21	-0.06
	IEP	0.31	0.75	+0.44
Accessibility		**2**	**4**	**+2**
Readability		11.7	5.4	-6.3
Word Count		10	8	-2

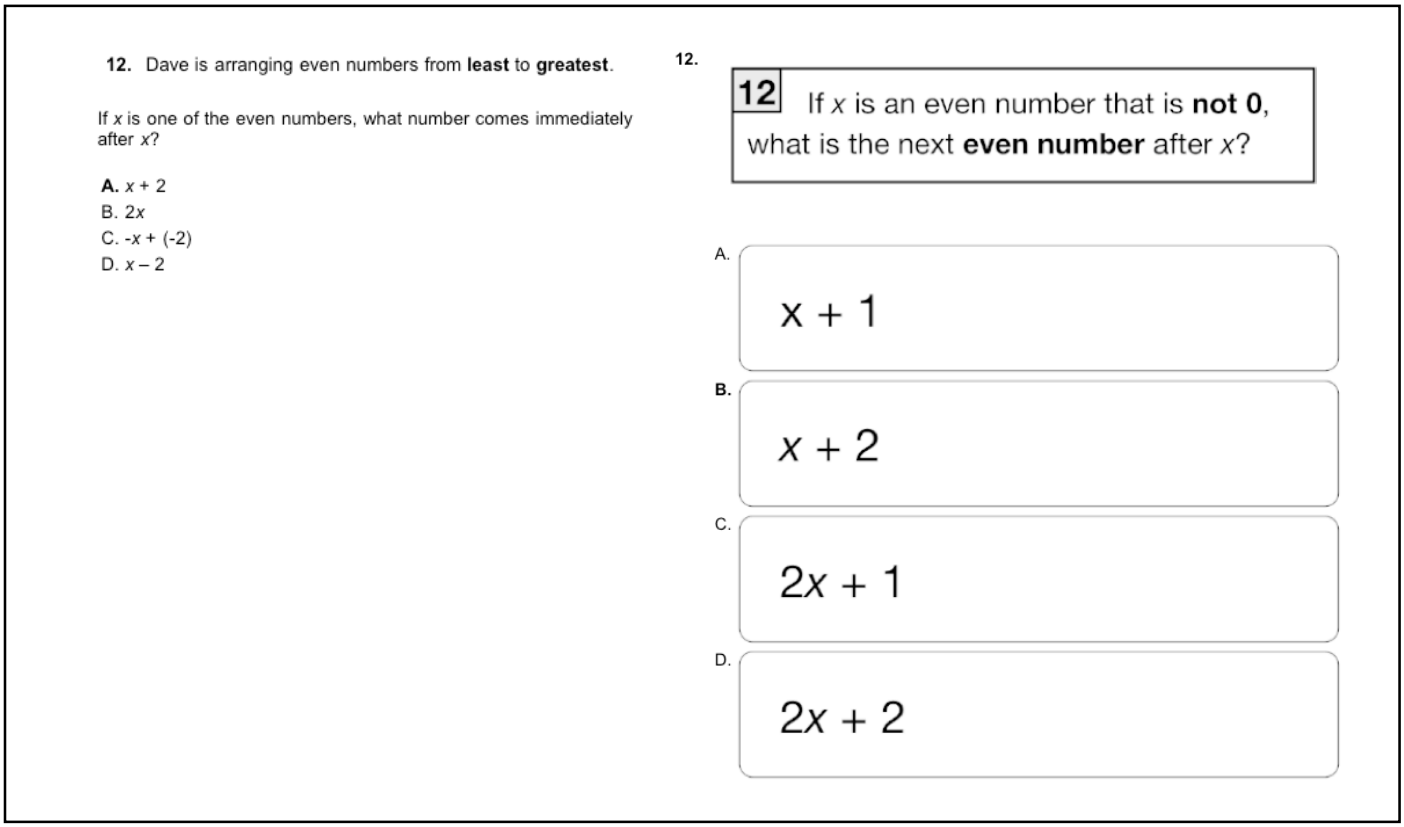


Figure B12. Item #12 in original and modified forms.

Table B12.
Item Statistics for Item #12

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.52**	**0.66**	**+0.14**
	No IEP	0.58	0.72	+0.14
	IEP	0.37	0.49	+0.12
Discrimination (*D*)	**Total**	**0.42**	**0.24**	**-0.19**
	No IEP	0.23	0.14	-0.09
	IEP	0.62	0.08	-0.53
Accessibility		**3**	**3**	**0**
Readability		7.3	3.9	-3.4
Word Count		23	18	-5

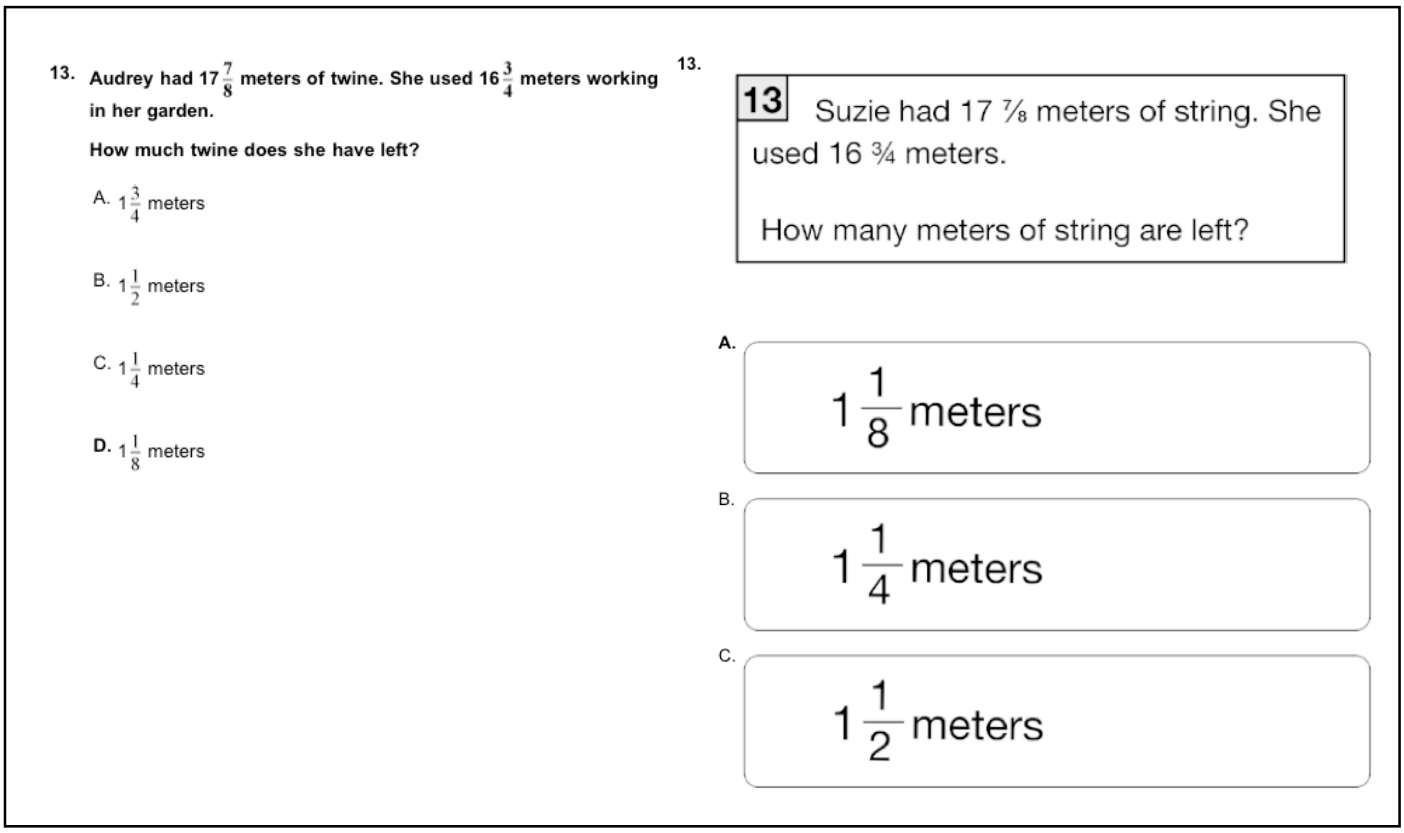


Figure B13. Item #13 in original and modified forms.

Table B13.
Item Statistics for Item #13

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.50**	**0.59**	**+0.09**
	No IEP	0.59	0.63	+0.04
	IEP	0.23	0.45	+0.22
Discrimination (*D*)	**Total**	**0.68**	**0.64**	**-0.04**
	No IEP	-0.68	0.69	+0.01
	IEP	0.38	0.25	-0.13
Accessibility		**3**	**4**	**+1**
Readability		3.3	1.4	-0.9
Word Count		19	19	0

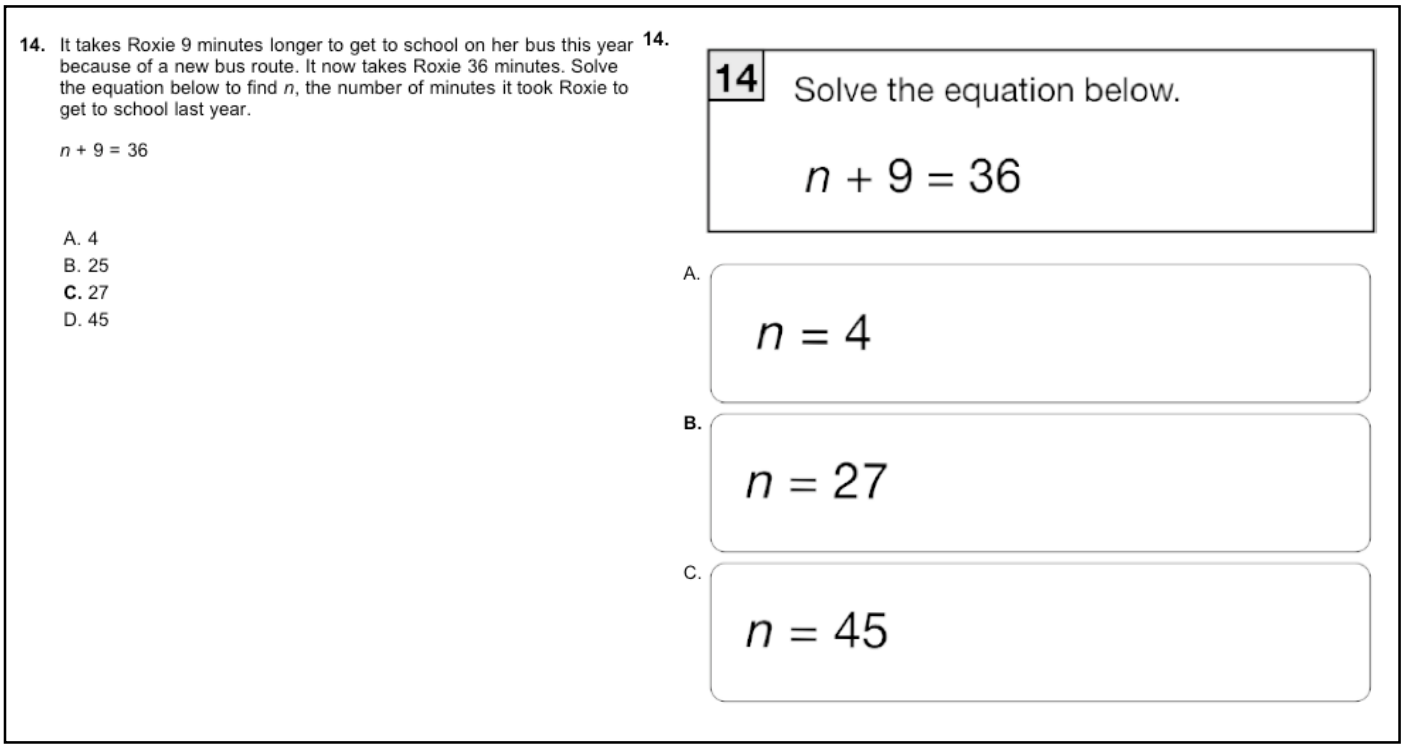


Figure B14. Item #14 in original and modified forms.

Table B14.
Item Statistics for Item #14

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.80**	**0.91**	**0.11**
	No IEP	0.84	0.96	+0.12
	IEP	0.67	0.73	+0.06
Discrimination (*D*)	**Total**	**0.32**	**0.27**	**-0.05**
	No IEP	0.16	0.10	-0.07
	IEP	0.62	0.50	-0.12
Accessibility		**2**	**4**	**+2**
Readability		5.3	9.5	+4.2
Word Count		47	4	-43

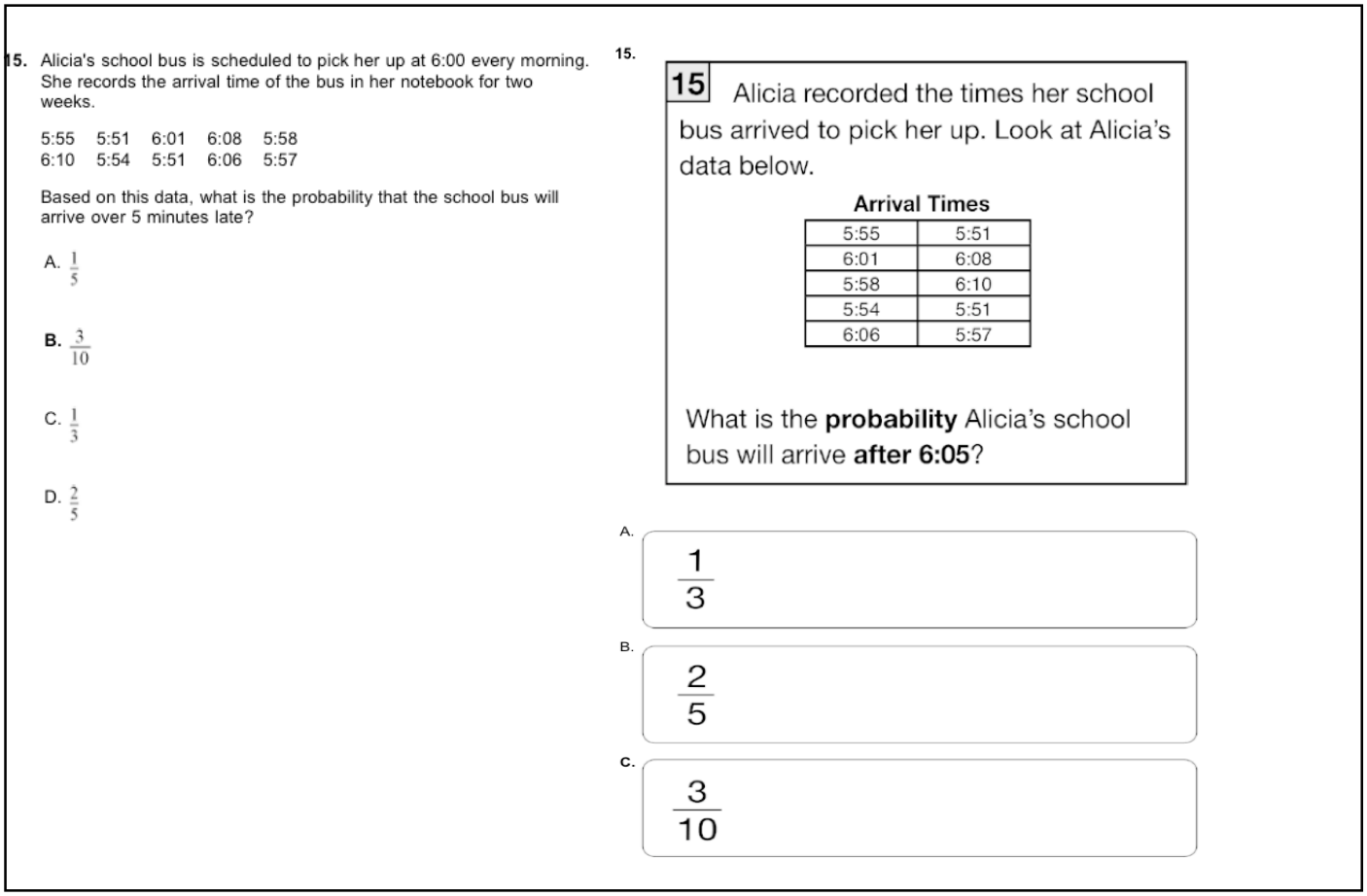


Figure B15. Item #15 in original and modified forms.

Table B15.
Item Statistics for Item #15

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.55**	**0.47**	**-0.08**
	No IEP	0.58	0.53	-0.05
	IEP	0.46	0.27	-0.19
Discrimination (*D*)	**Total**	**0.33**	**0.47**	**+0.15**
	No IEP	0.27	0.33	+0.06
	IEP	0.38	0.33	-0.05
Accessibility		**3**	**3**	**0**
Readability		7.5	6.6	-0.9
Word Count		46	29	-17

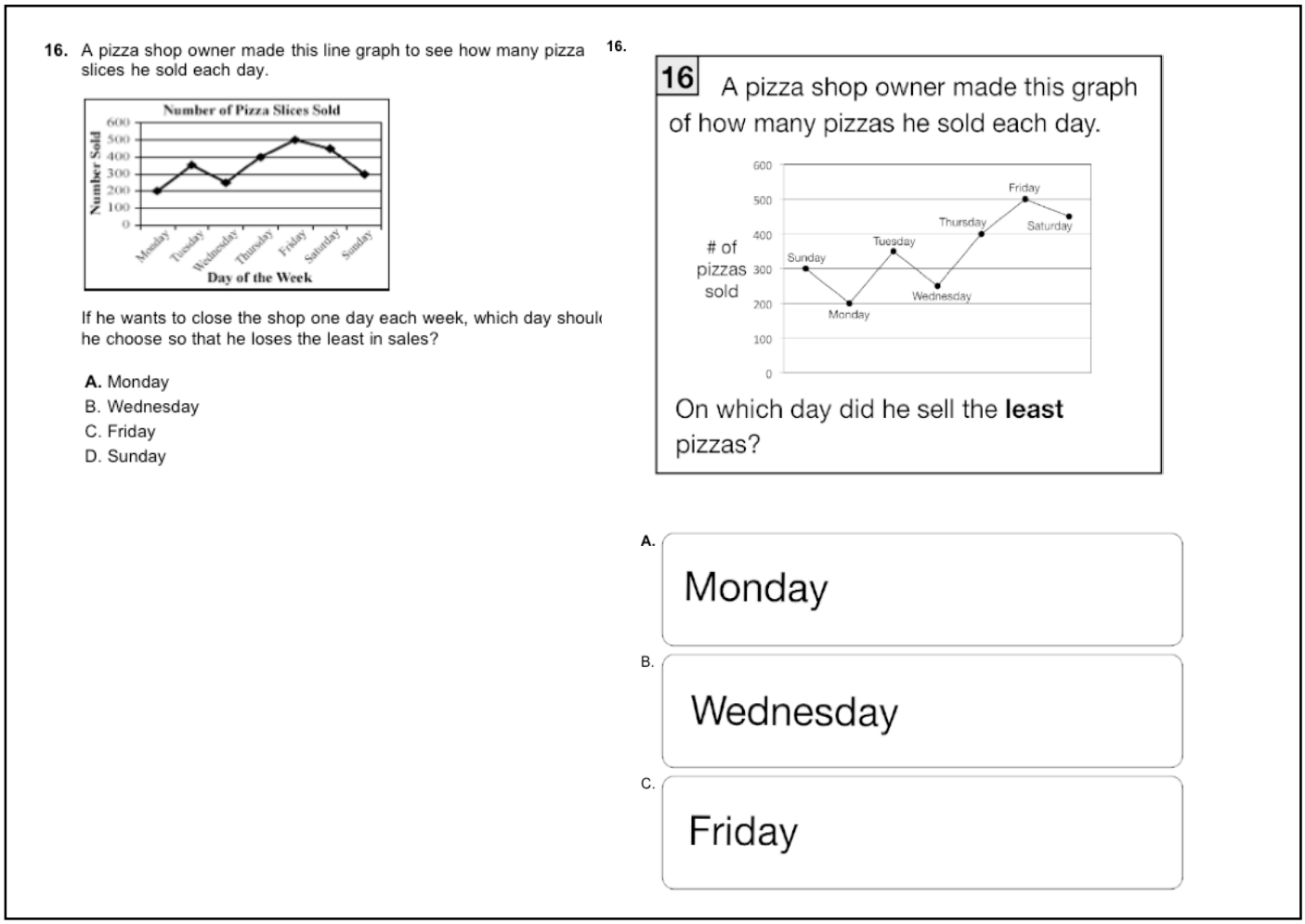


Figure B16. Item #16 in original and modified forms.

Table B16.
Item Statistics for Item #16

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.77**	**0.96**	**+0.19**
	No IEP	0.82	0.99	+0.17
	IEP	0.62	0.86	+0.24
Discrimination (*D*)	**Total**	**0.44**	**0.13**	**-0.31**
	No IEP	0.16	0.02	-0.14
	IEP	0.46	0.33	-0.13
Accessibility		**2**	**3**	**+1**
Readability		5.6	3.6	-2.0
Word Count		42	24	-18

17. Theo has 3 dogs: Tucker, Max, and Pippie. He can only walk 2 dogs at a time.

How many different combinations of 2 dogs can Theo pick to take for a walk out of Tucker, Max, and Pippie?

A. 1
B. 3
C. 12
D. 18

17.

17 Tom has 3 dogs: Tucker, Max, and Pippie.


Tucker


Max


Pippie

How many different combinations of 2 dogs can Tom walk?

A. 1

B. 3

C. 6

Figure B17. Item #17 in original and modified forms.

Table B17.
Item Statistics for Item #17

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.58**	**0.45**	**-0.13**
	No IEP	0.59	0.44	-0.15
	IEP	0.52	0.45	-0.07
Discrimination (*D*)	**Total**	**0.33**	**-0.20**	**-0.13**
	No IEP	0.28	0.29	+0.01
	IEP	0.62	0.33	-0.28
Accessibility		**3**	**3**	**0**
Readability		4.7	5.7	+1.0
Word Count		38	18	-20
Student Variables				
Comprehension [1]	**Total**	**4.4 (1.4)**	**5.1 (1.3)**	**+0.7**
	No IEP	4.5 (1.5)	5.3 (1.2)	+0.8
	IEP	4.1 (1.3)	4.4 (1.5)	+0.3
Cognitive Demand [1]	**Total**	**2.7 (1.4)**	**2.1 (1.6)**	**-0.6**
	No IEP	2.5 (1.3)	1.9 (1.5)	-0.6
	IEP	3.2 (1.7)	2.9 (1.4)	-0.3
OTL [1]	**Total**	**3.4 (1.6)**	**4.1 (1.5)**	**+0.7**
	No IEP	3.4 (1.5)	4.3 (1.5)	+0.9
	IEP	3.2 (1.8)	3.4 (1.5)	+0.2
Confidence [1]	**Total**	**4.0 (1.7)**	**4.8 (1.5)**	**+0.8**
	No IEP	4.0 (1.8)	5.2 (1.3)	+5.2
	IEP	3.9 (1.6)	3.7 (1.7)	-0.2
Help [1]	**Total**	**2.1 (1.5)**	**3.6 (1.4)**	**+1.5**
	No IEP	2.2 (1.7)	3.3 (2.5)	+1.1
	IEP	1.7 (1.2)	5.0 (1.4)	+3.3

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

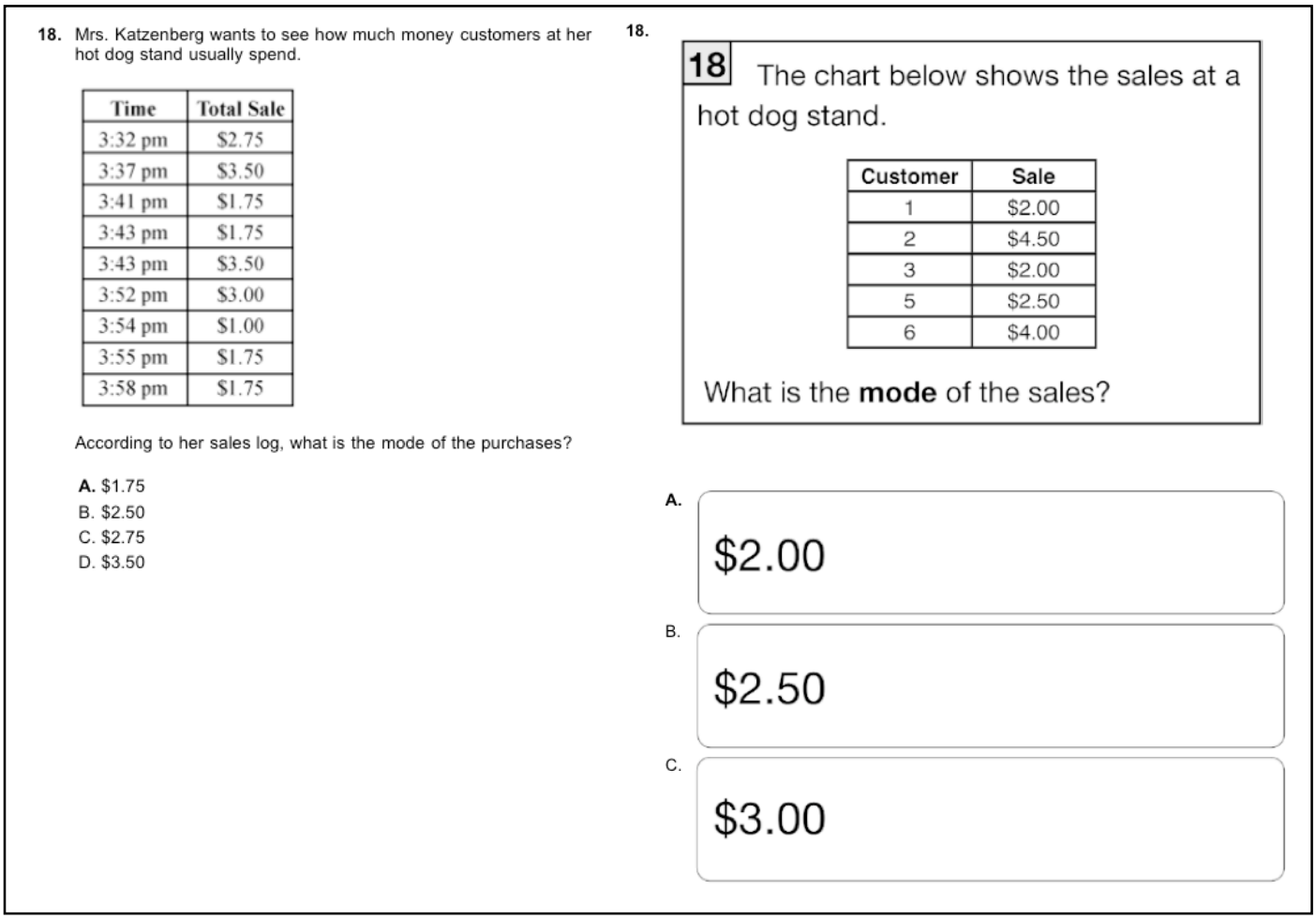


Figure B18. Item #18 in original and modified forms.

Table B18.
Item Statistics for Item #18

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.74**	**0.70**	**-0.04**
	No IEP	0.80	0.73	-0.07
	IEP	0.56	0.59	+0.03
Discrimination (*D*)	**Total**	**0.51**	**0.22**	**-0.29**
	No IEP	0.32	0.26	-0.06
	IEP	0.69	0.08	-0.61
Accessibility		**3**	**3**	**0**
Readability		6.8	2.0	-4.8
Word Count		28	18	-10

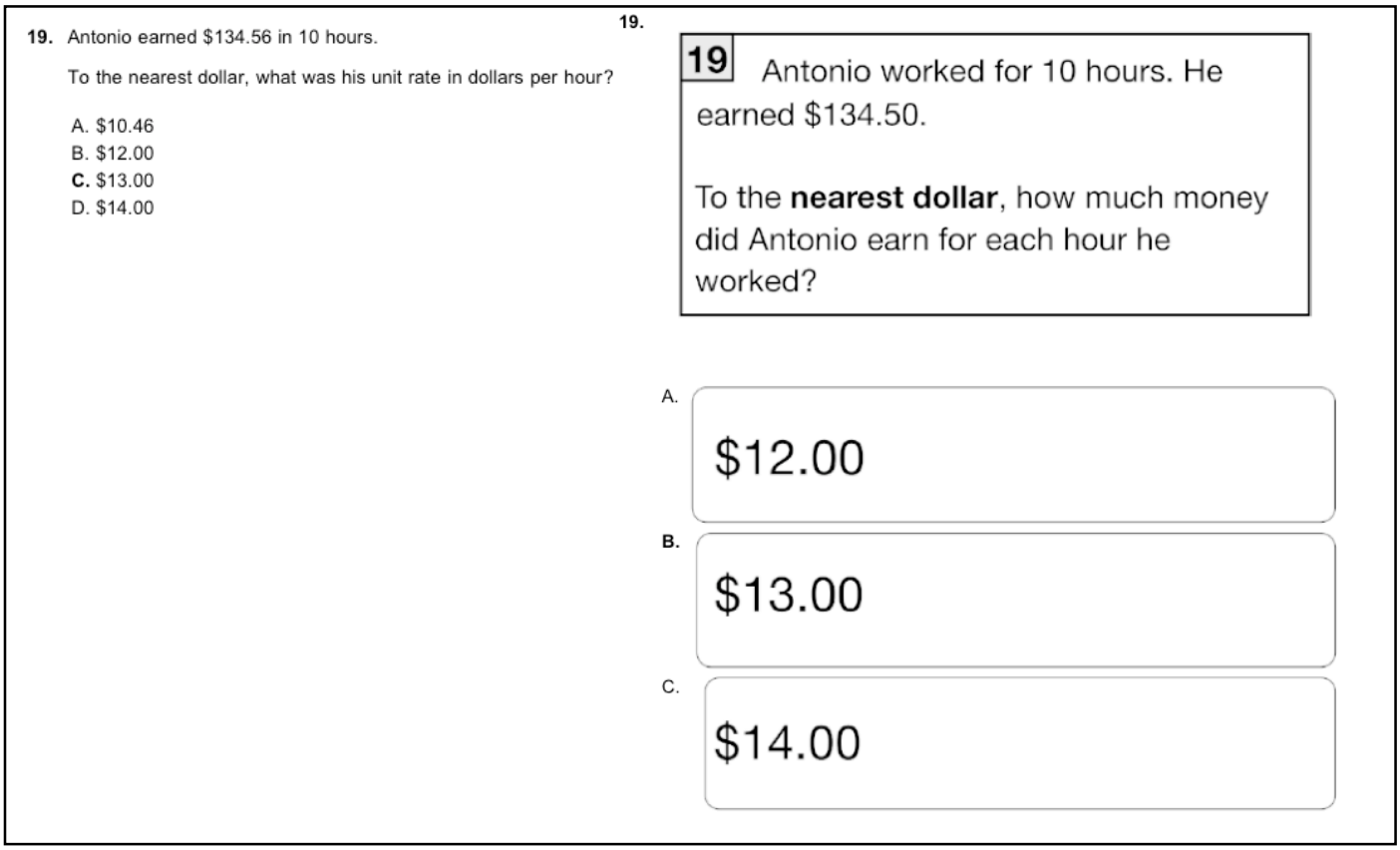


Figure B19. Item #19 in original and modified forms.

Table B19.
Item Statistics for Item #19

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.47**	**0.64**	**+0.17**
	No IEP	0.53	0.67	+0.14
	IEP	0.29	0.51	+0.22
Discrimination (*D*)	**Total**	**0.44**	**0.27**	**-0.17**
	No IEP	0.23	0.26	+0.03
	IEP	0.23	0.25	+0.02
Accessibility		**3**	**4**	**+1**
Readability		2.7	3.2	+0.5
Word Count		20	24	+4

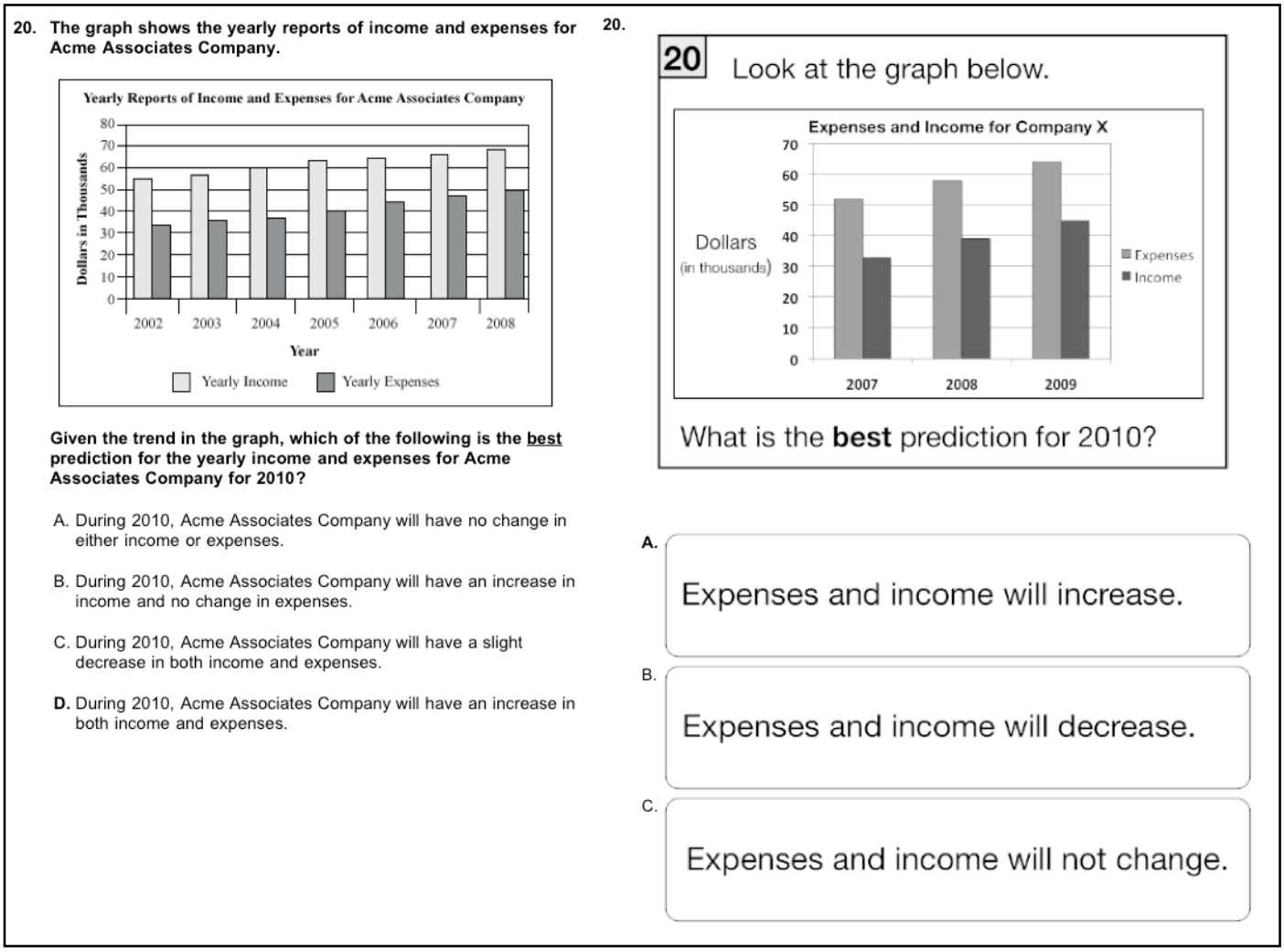


Figure B20. Item #20 in original and modified forms.

Table B20.
Item Statistics for Item #20

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.59**	**0.89**	**0.30**
	No IEP	0.64	0.93	+0.29
	IEP	0.42	0.76	+0.34
Discrimination (*D*)	**Total**	**0.57**	**0.27**	**-0.30**
	No IEP	0.37	0.19	-0.18
	IEP	-0.54	0.33	-0.21
Accessibility		**2**	**3**	**+1**
Readability		10.6	6.7	-3.9
Word Count		99	28	71

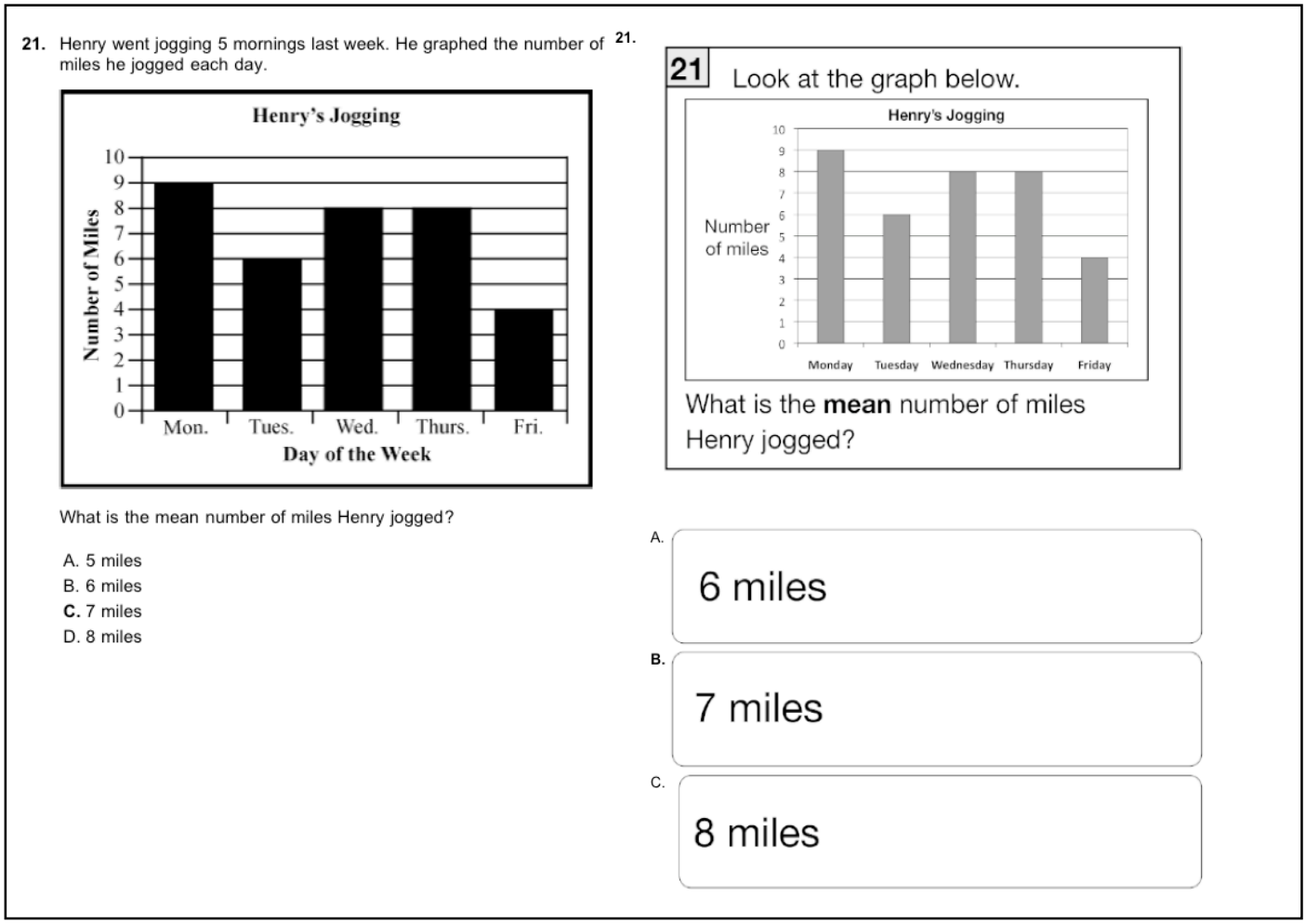


Figure B21. Item #21 in original and modified forms.

Table B21.
Item Statistics for Item #21

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.53**	**0.64**	**+0.11**
	No IEP	0.59	0.73	+0.14
	IEP	0.35	0.31	-0.04
Discrimination (*D*)	**Total**	**0.61**	**0.69**	**+0.08**
	No IEP	0.46	0.52	+0.07
	IEP	0.46	0.42	-0.04
Accessibility		**3**	**4**	**+1**
Readability		3.8	2.9	-0.9
Word Count		26	14	-12

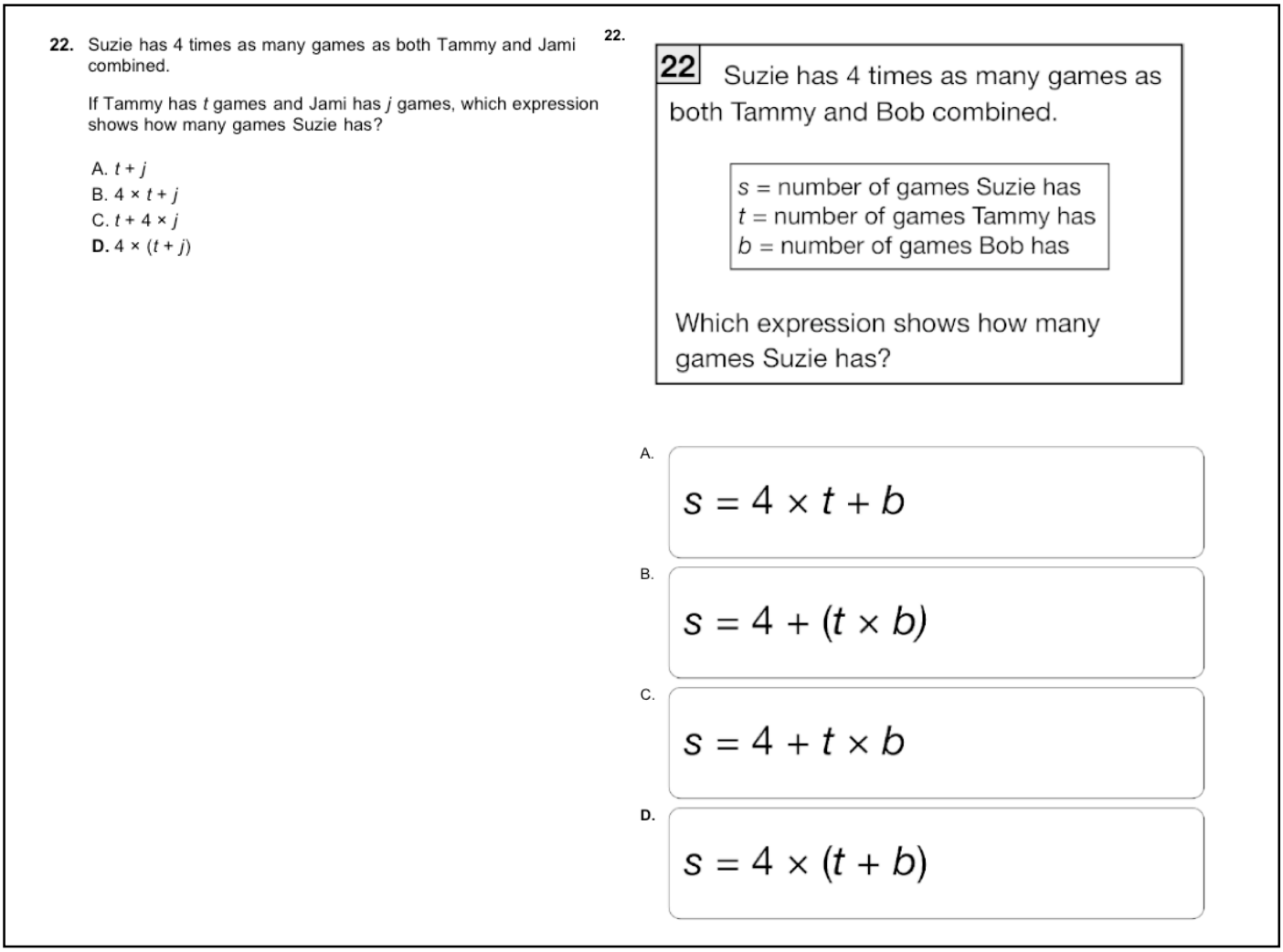


Figure B22. Item #22 in original and modified forms.

Table B22.
Item Statistics for Item #22

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.54**	**0.52**	**-0.02**
	No IEP	0.62	0.57	-0.05
	IEP	0.29	0.33	+0.04
Discrimination (*D*)	**Total**	**0.72**	**0.55**	**-0.17**
	No IEP	0.46	0.48	+0.02
	IEP	0.62	0.25	-0.37
Accessibility		**3**	**3**	**0**
Readability		6.1	6.7	+0.6
Word Count		31	42	+11

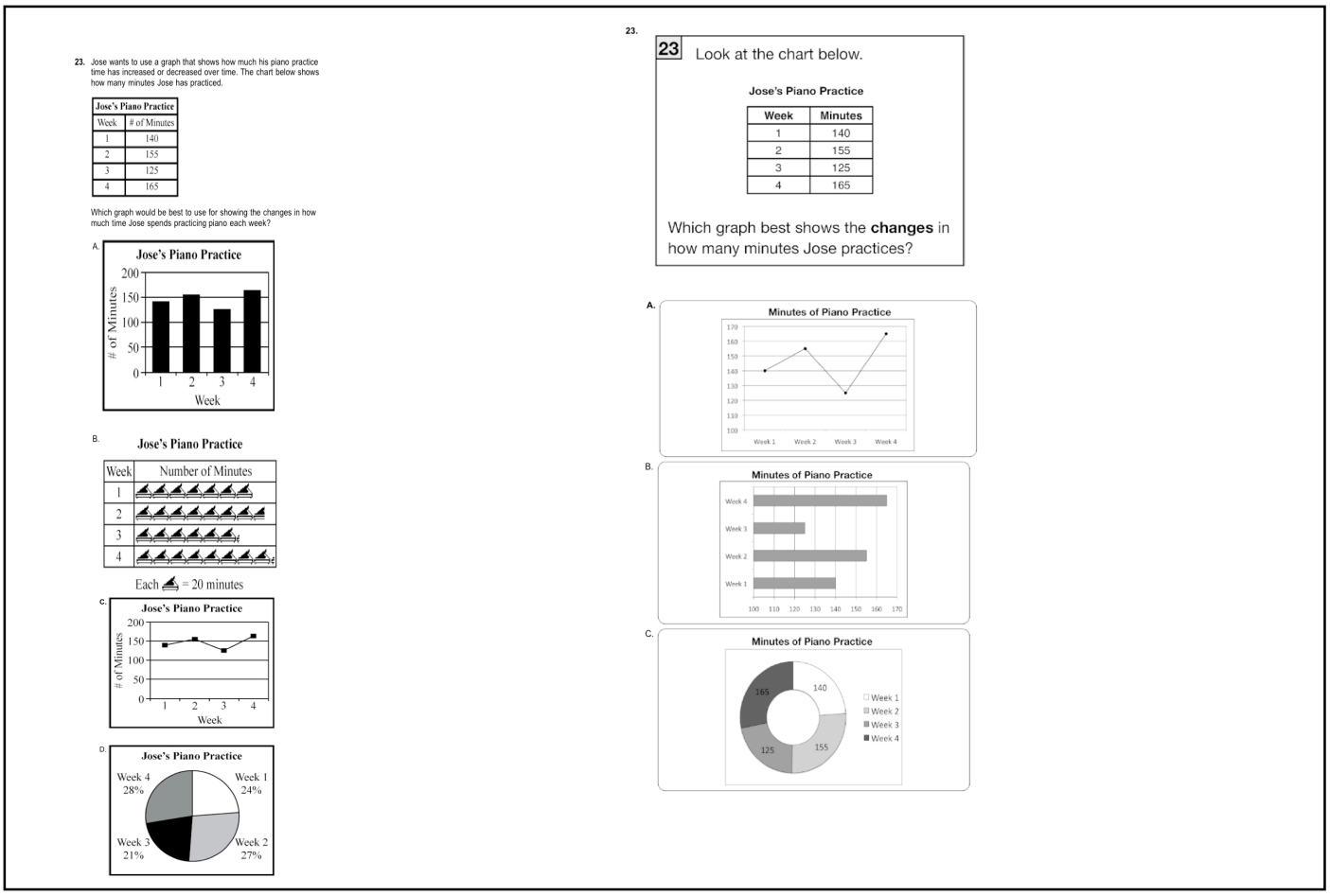


Figure B23. Item #23 in original and modified forms.

Table B23.
Item Statistics for Item #23

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.30**	**0.59**	**+0.29**
	No IEP	0.29	0.64	+0.35
	IEP	0.31	0.43	+0.12
Discrimination (*D*)	**Total**	**0.22**	**0.65**	**+0.44**
	No IEP	0.38	0.50	+0.12
	IEP	0.15	0.67	+0.51
Accessibility		**2**	**3**	**+1**
Readability		7.9	4.4	
Word Count		51	17	

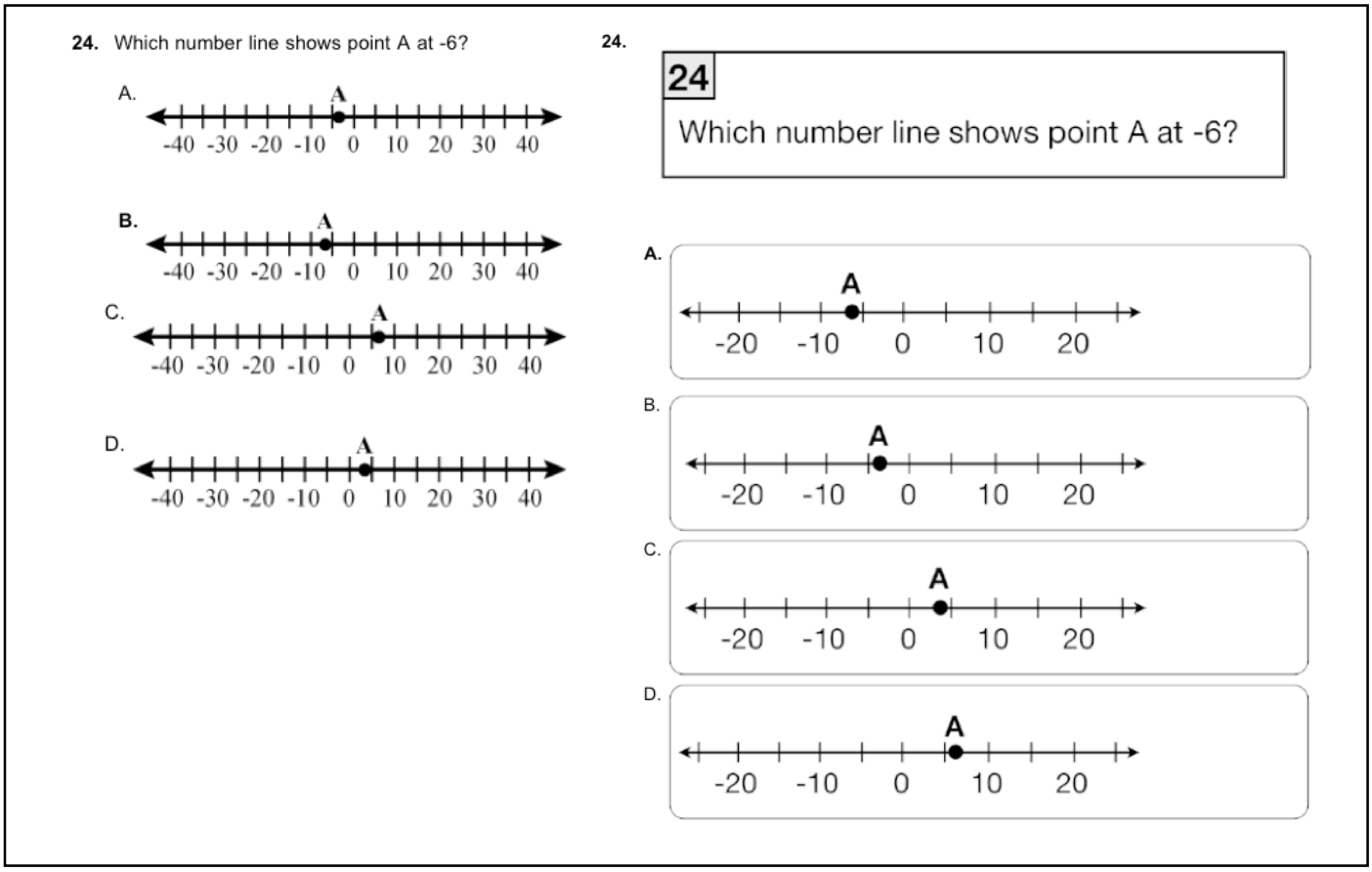


Figure B24. Item #24 in original and modified forms.

Table B24.
Item Statistics for Item #24

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.65**	**0.82**	**+0.17**
	No IEP	0.71	0.88	+0.17
	IEP	0.44	0.63	+0.19
Discrimination (*D*)	**Total**	**0.40**	**0.29**	**-0.11**
	No IEP	0.15	0.07	-0.07
	IEP	0.31	0.42	+0.11
Accessibility		**2**	**3**	**+1**
Readability		1.9	1.9	0.0
Word Count		8	8	0

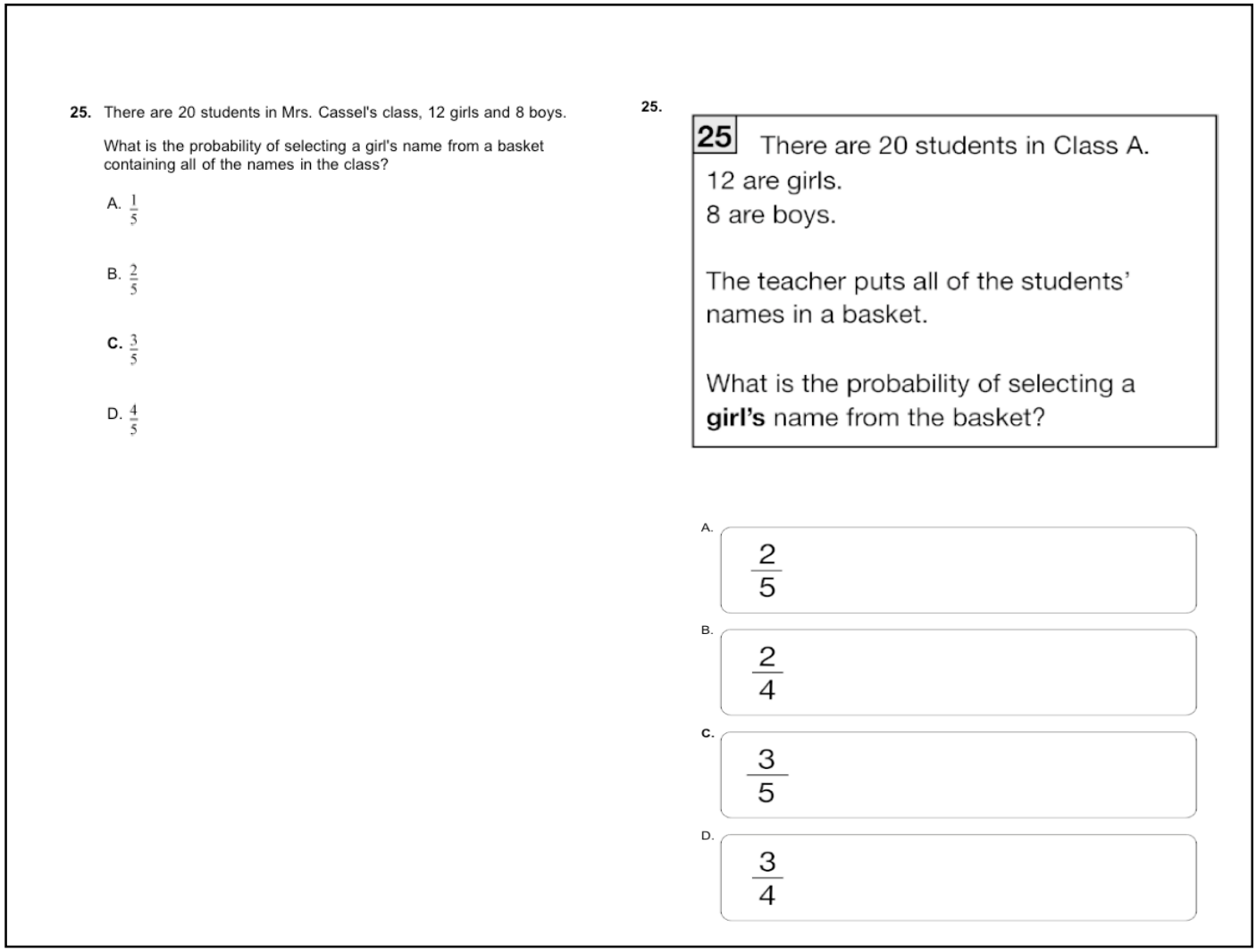


Figure B25. Item #25 in original and modified forms.

Table B25.
Item Statistics for Item #25

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.58**	**0.48**	**-0.10**
	No IEP	0.63	0.53	-0.10
	IEP	0.44	0.29	-0.15
Discrimination (*D*)	**Total**	**0.57**	**0.60**	**+0.03**
	No IEP	0.57	0.55	-0.02
	IEP	0.23	0.42	+0.19
Accessibility		**3**	**4**	**+1**
Readability		5.7	3.7	-2.0
Word Count		33	36	+3

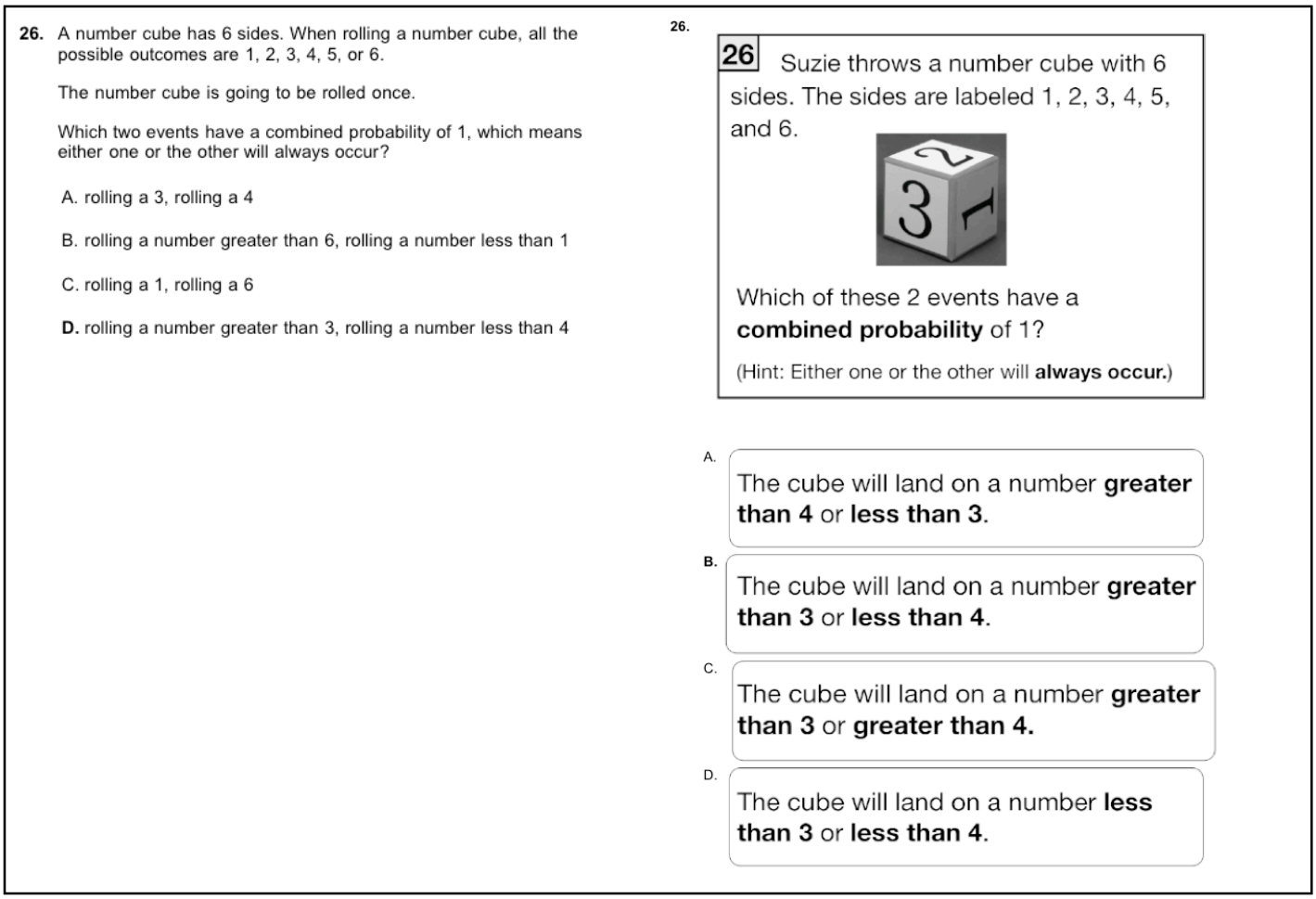


Figure B26. Item #26 in original and modified forms.

Table B26.
Item Statistics for Item #26

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.37**	**0.30**	**-0.07**
	No IEP	0.44	0.27	-0.17
	IEP	0.17	0.41	+0.24
Discrimination (*D*)	Total	**0.35**	**0.02**	**-0.33**
	No IEP	0.16	0.05	-0.11
	IEP	0.23	0.08	-0.15
Accessibility		**3**	**3**	**0**
Readability		5.8	3.0	-2.8
Word Count		86	95	+9

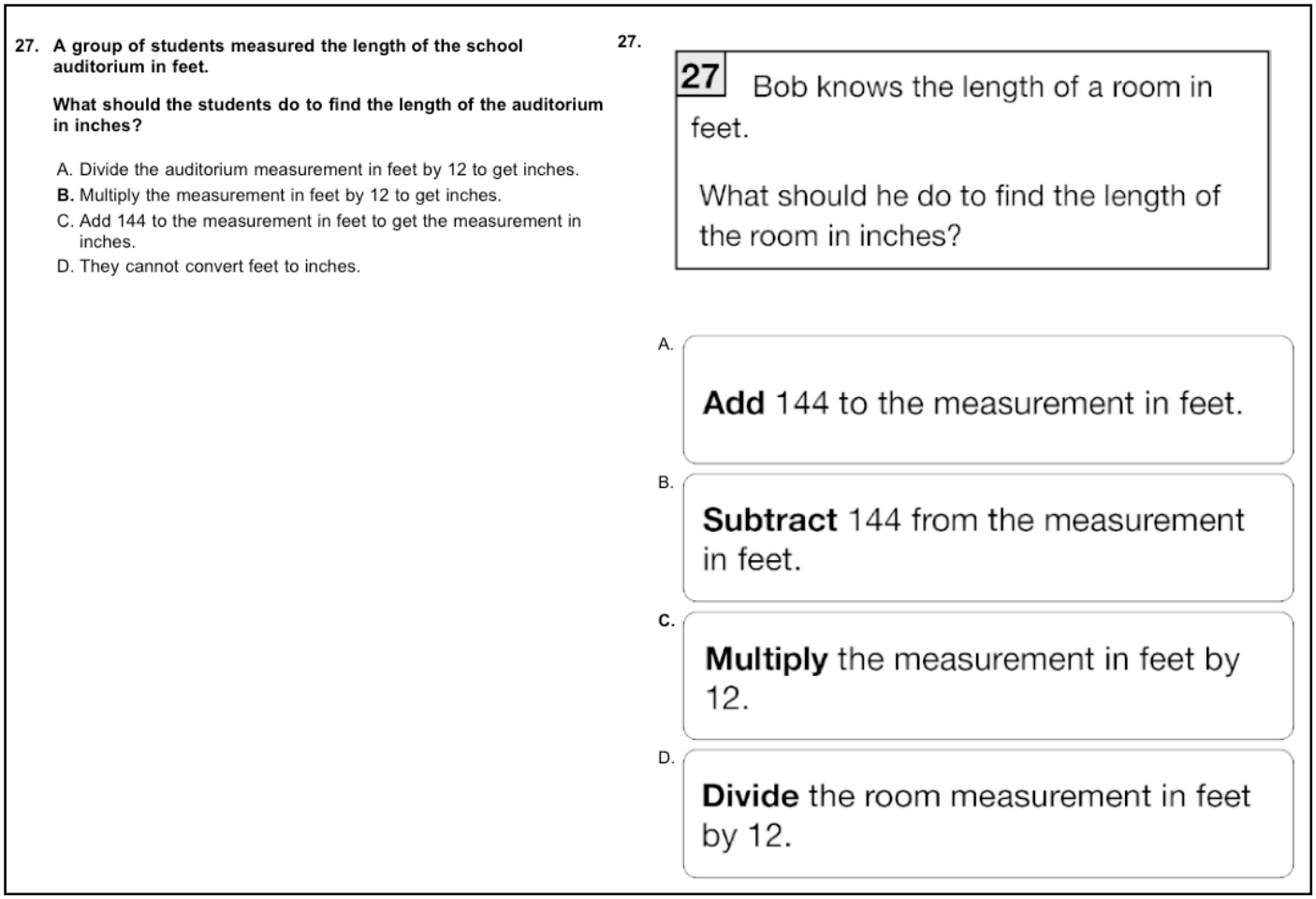


Figure B27. Item #27 in original and modified forms.

Table B27.
Item Statistics for Item #10

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.33**	**0.43**	**+0.10**
	No IEP	0.34	0.45	+0.12
	IEP	0.29	0.35	+0.06
Discrimination (*D*)	**Total**	**0.42**	**0.22**	**-0.21**
	No IEP	0.35	0.19	-0.16
	IEP	-0.08	0.33	+0.41
Accessibility		**3**	**3**	**0**
Readability		7.3	2.3	-5.0
Word Count		67	22	-45

28. Martha tracked the number of visitors to the water park and the average temperature in the table below.

Water Park Attendance	
Temperature (°F)	Number of Visitors
100	800
95	725
90	700
85	650
80	575
75	490

What does this data show about the relationship between the number of visitors and temperature?

A. The number of visitors to the water park increases as the temperature increases.
B. The number of visitors to the water park increases as the temperature decreases.
C. The number of visitors to the water park decreases as the temperature increases.
D. There is no relationship between the number of visitors and the temperature.

28.

28 Suzie tracked the temperature and the number of visitors to the water park.

Water Park Attendance

Temperature	Number of Visitors
100	800
95	725
90	700
85	575

What does the table show?

A. The number of visitors **increases** when the temperature **increases**.

B. The number of visitors **increases** when the temperature **decreases**.

C. There is **no relation** between the number of visitors and the temperature.

Figure B28. Item #28 in original and modified forms.

Table B28.
Item Statistics for Item #28

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.59**	**0.71**	**+0.12**
	No IEP	0.65	0.79	+0.14
	IEP	0.40	0.43	+0.03
Discrimination (*D*)	**Total**	**0.72**	**0.64**	**-0.08**
	No IEP	0.46	0.43	-0.03
	IEP	0.62	0.67	+0.05
Accessibility		**3**	**4**	**+1**
Readability		11.1	10.1	-1.0
Word Count		84	48	-36
Student Variables				
Comprehension [1]	**Total**	**4.6 (1.8)**	**4.8 (1.7)**	**+0.2**
	No IEP	4.7 (1.8)	5.0 (1.6)	+0.3
	IEP	4.2 (1.8)	4.1 (1.7)	-0.1
Cognitive Demand [1]	**Total**	**2.7 (1.4)**	**2.3 (1.6)**	**-0.4**
	No IEP	2.6 (1.7)	2.0 (1.5)	-0.6
	IEP	3.4 (1.8)	3.3 (1.6)	-0.1
OTL [1]	**Total**	**3.6 (1.7)**	**4.0 (1.6)**	**+0.4**
	No IEP	3.7 (1.7)	4.1 (1.5)	+0.4
	IEP	3.2 (1.8)	3.9 (1.7)	+0.7
Confidence [1]	**Total**	**4.3 (1.7)**	**4.7 (1.6)**	**+0.4**
	No IEP	4.4 (1.7)	5.0 (1.5)	+0.6
	IEP	4.2 (1.7)	3.8 (1.7)	-0.6
Help [1]	**Total**	**2.1 (1.6)**	**3.5 (2.1)**	**+1.4**
	No IEP	2.1 (1.7)	3.7 (2.2)	+1.6
	IEP	2.0 (1.7)	3.3 (2.2)	+1.3

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

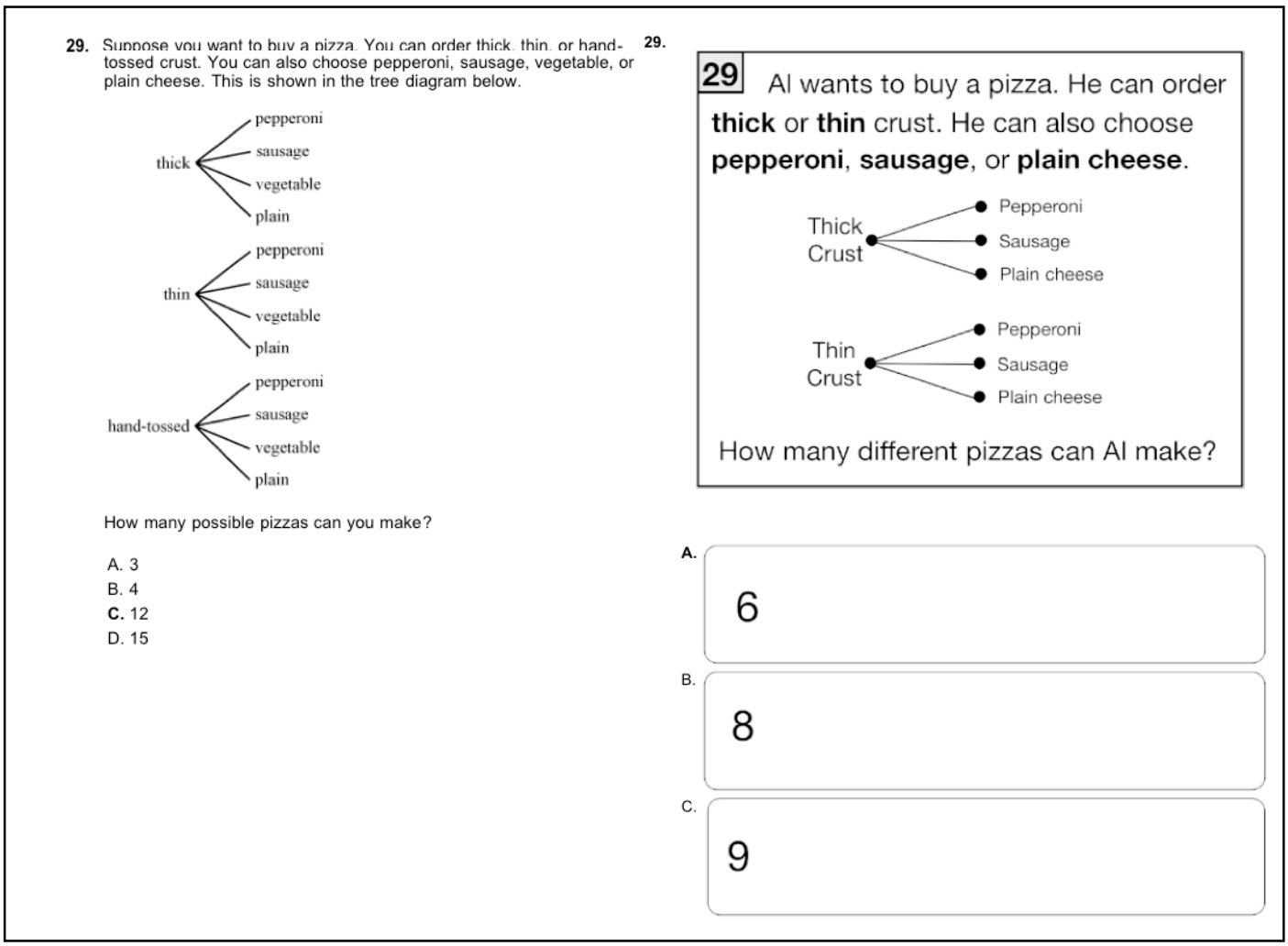


Figure B29. Item #6 in original and modified forms.

Table B29.
Item Statistics for Item #29

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.66**	**0.64**	**-0.02**
	No IEP	0.68	0.67	-0.01
	IEP	0.60	0.51	-0.09
Discrimination (*D*)	**Total**	**0.55**	**0.65**	**+0.11**
	No IEP	0.57	0.67	+0.10
	IEP	0.77	0.42	-0.35
Accessibility		**2**	**4**	**+2**
Readability		6.0	7.2	+1.2
Word Count		41	41	0

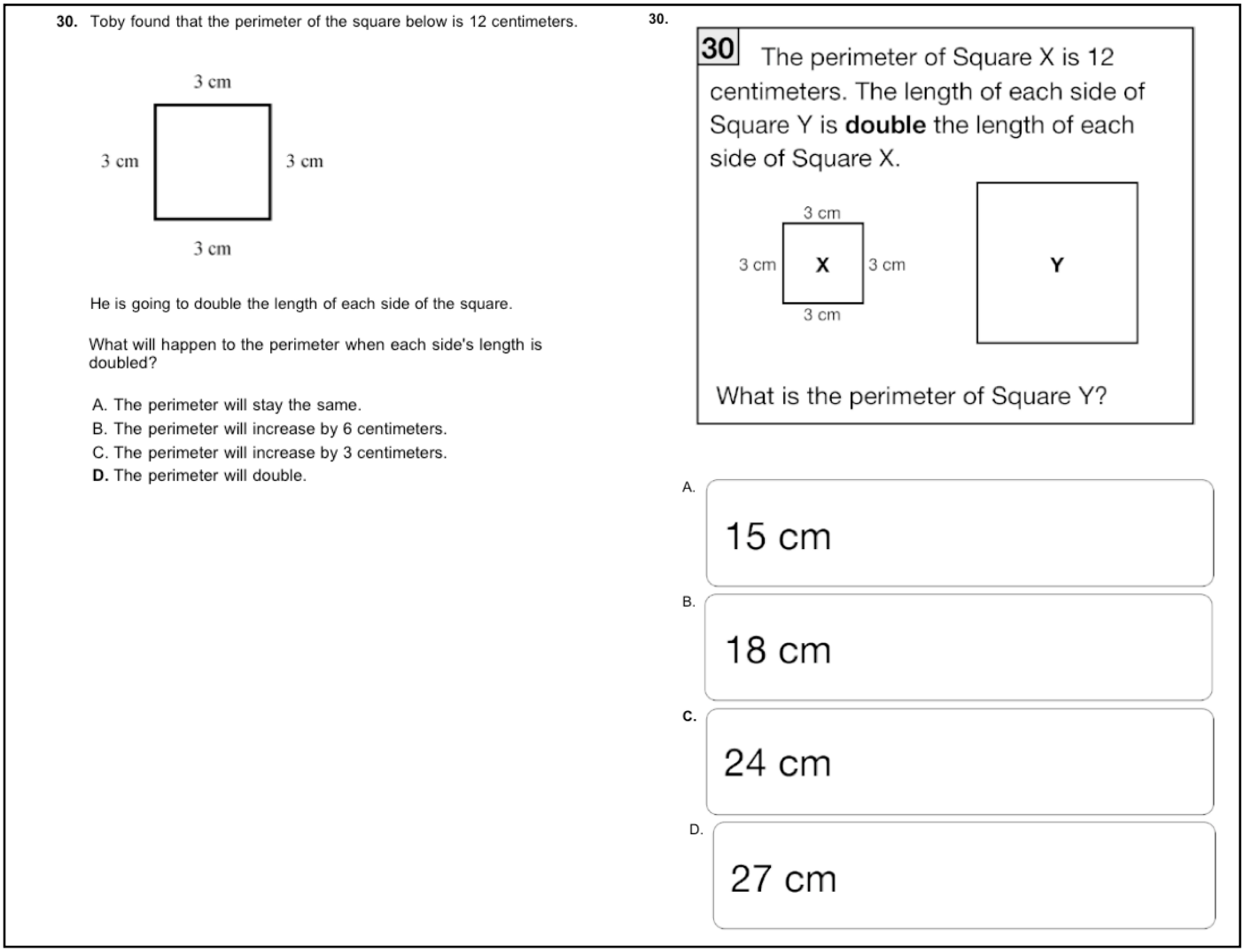


Figure B30. Item #30 in original and modified forms.

Table B30.
Item Statistics for Item #30

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.51**	**0.75**	**+0.24**
	No IEP	0.58	0.83	+0.25
	IEP	0.33	0.49	+0.16
Discrimination (*D*)	**Total**	**0.77**	**0.58**	**-0.19**
	No IEP	0.47	0.40	-0.06
	IEP	0.62	0.33	-0.28
Accessibility		**3**	**3**	**0**
Readability		7.6	5.3	-2.3
Word Count		61	33	-28

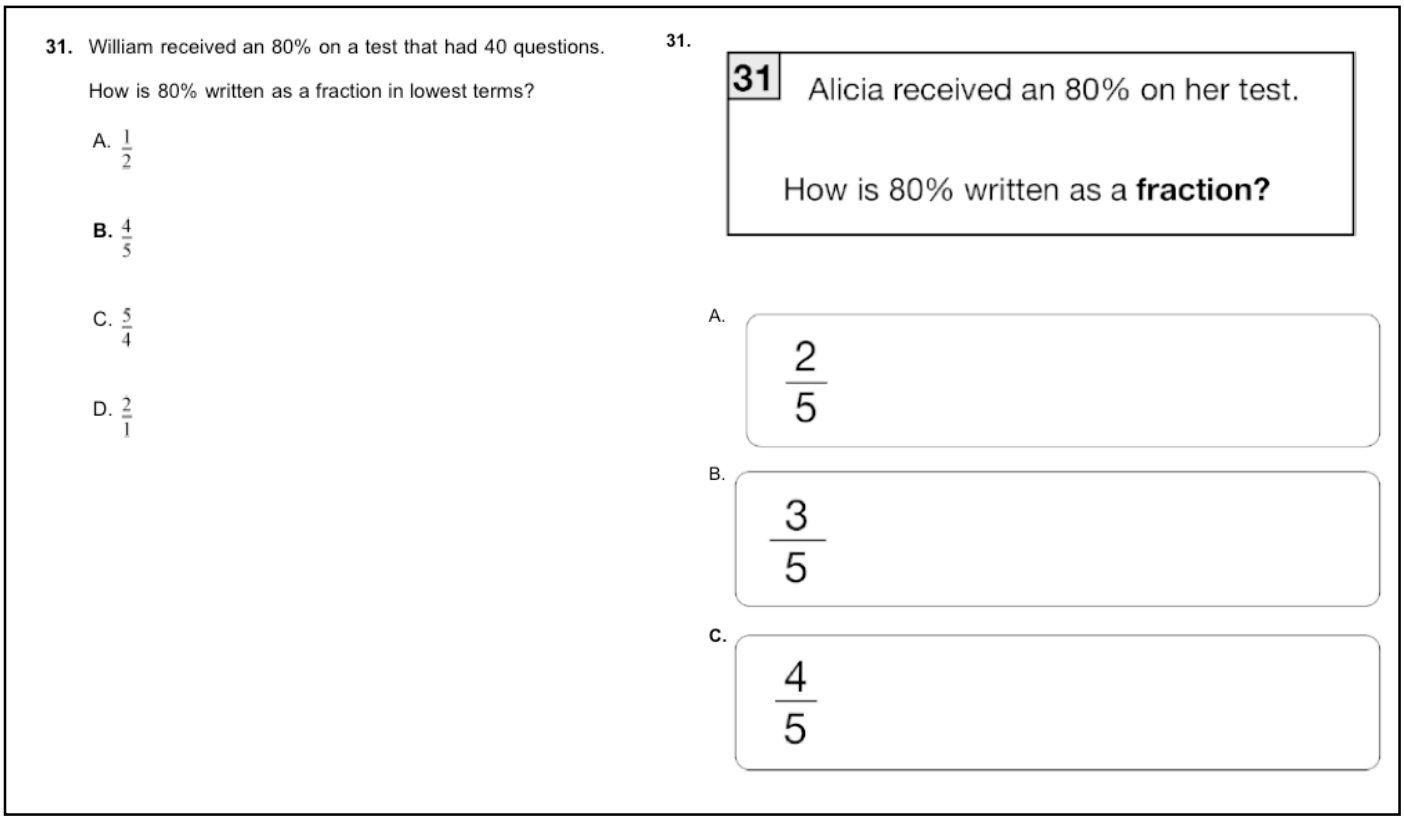


Figure B31. Item #31 in original and modified forms.

Table B31.
Item Statistics for Item #31

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.75**	**0.62**	**-0.13**
	No IEP	0.79	0.69	-0.10
	IEP	0.60	0.41	-0.19
Discrimination (*D*)	**Total**	**0.47**	**0.60**	**+0.13**
	No IEP	0.50	0.03	+0.31
	IEP	0.31	0.42	+0.11
Accessibility		**3**	**4**	**+1**
Readability		3.8	2.8	-1.0
Word Count		21	14	-7

32. Craig is going to roll a number cube labeled with numbers 1 through 6, like the one below.

What is the probability that he will roll a 1, 2, or 3?

A. 25%
B. 33%
C. 40%
D. 50%

32.

32 Roy throws a number cube with 6 sides. The sides are labeled 1, 2, 3, 4, 5, and 6.

What is the **probability** the number cube will land on 1, 2, or 3?

A. 25%

B. 33%

C. 50%

Figure B32. Item #32 in original and modified forms.

Table B32.
Item Statistics for Item #32

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.43**	**0.61**	**+0.18**
	No IEP	0.49	0.70	+0.21
	IEP	0.27	0.31	+0.04
Discrimination (*D*)	**Total**	**0.65**	**0.69**	**+0.05**
	No IEP	0.57	0.55	-0.02
	IEP	0.23	0.25	+0.02
Accessibility		**3**	**4**	**+1**
Readability		4.9	2.2	-2.7
Word Count		31	33	+2

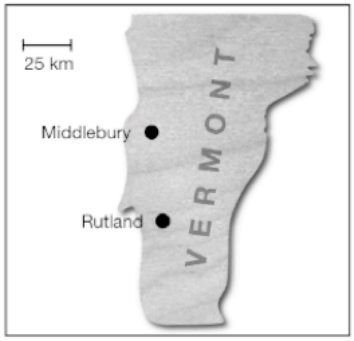
33. Look at the map below.

About how far is Middlebury from Rutland?

A. 25 kilometers
B. 50 kilometers
C. 50 miles
D. 100 miles

33.

33 Look at the map below.



About how far is Middlebury from Rutland?

A. 25 km

B. 50 km

C. 100 km

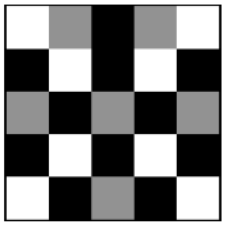
Figure B33. Item #33 in original and modified forms.

Table B33.
Item Statistics for Item #33

Item Variables		Original	Modified	Δ
Difficulty (*p*)	Total	**0.69**	**0.72**	**0.03**
	No IEP	0.73	0.75	+0.02
	IEP	0.56	0.63	+0.10
Discrimination (*D*)	**Total**	**0.53**	**0.29**	**-0.24**
	No IEP	0.30	0.24	-0.06
	IEP	0.31	0.50	+0.19
Accessibility		**2**	**4**	**+2**
Readability		4.4	4.4	0.0
Word Count		12	12	0
Student Variables				
Comprehension [1]	**Total**	**4.7 (1.7)**	**4.9 (1.2)**	**+0.2**
	No IEP	4.7 (1.7)	5.1 (1.2)	+0.4
	IEP	4.7 (1.5)	4.5 (1.4)	-0.2
Cognitive Demand [1]	**Total**	**2.7 (1.6)**	**2.1 (1.1)**	**-0.6**
	No IEP	2.5 (1.5)	2.1 (1.2)	-0.4
	IEP	3.2 (2.0)	2.3 (1.0)	-0.9
OTL [1]	**Total**	**3.5 (1.6)**	**3.0 (1.6)**	**-0.5**
	No IEP	3.5 (1.6)	3.0 (1.6)	-0.5
	IEP	3.6 (1.6)	3.1 (1.6)	-0.5
Confidence [1]	**Total**	**4.5 (1.6)**	**4.5 (1.5)**	**0.0**
	No IEP	4.4 (1.7)	4.8 (1.3)	+0.4
	IEP	4.8 (1.4)	3.6 (1.7)	-1.2
Help [1]	**Total**	**2.0 (1.6)**	**3.3 (2.1)**	**+1.3**
	No IEP	2.0 (1.6)	3.3 (2.2)	+1.3
	IEP	2.0 (1.7)	3.3 (2.1)	+1.3

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

34. Christian is going to drop a pebble on the square below.



If the pebble lands on one color, what is the probability that it will land on a gray area?

A. $\frac{6}{25}$

34.

34 There are three different colors of marbles in a bag. Look at the chart below.

Color	Number of Marbles
Red	8
Green	6
Blue	11

Jamie reaches into the bag and removes a marble. What is the probability the marble is **green**?

A. $\frac{6}{19}$

B. $\frac{6}{25}$

C. $\frac{19}{25}$

Figure B34. Item #34 in original and modified forms.

Table B34.
Item Statistics for Item #34

Item Variables		Original	Modified	Δ
Difficulty (*p*)	**Total**	**0.70**	**0.75**	**0.05**
	No IEP	0.76	0.79	+0.03
	IEP	0.54	0.63	+0.09
Discrimination (*D*)	**Total**	**0.49**	**0.51**	**+0.02**
	No IEP	0.43	0.43	+0.00
	IEP	0.08	0.58	+0.51
Accessibility		**3**	**4**	**+1**
Readability		5.9	5.3	-0.6
Word Count		30	32	-2
Student Variables				
Comprehension [1]	**Total**	**4.8 (1.3)**	**4.6 (1.7)**	**-0.2**
	No IEP	4.9 (1.3)	4.7 (1.7)	-0.2
	IEP	4.4 (1.4)	4.3 (1.7)	-0.1
Cognitive Demand [1]	**Total**	**2.5 (1.3)**	**2.8 (1.6)**	**+0.3**
	No IEP	2.5 (1.4)	2.7 (1.7)	+0.2
	IEP	2.5 (1.3)	3.1 (1.4)	+0.6
OTL [1]	**Total**	**3.5 (1.5)**	**4.1 (1.5)**	**+0.6**
	No IEP	3.6 (1.4)	4.2 (1.5)	+0.6
	IEP	3.3 (1.9)	4.1 (1.7)	+0.8
Confidence [1]	**Total**	**4.3 (1.7)**	**4.2 (1.8)**	**-0.1**
	No IEP	4.5 (1.7)	4.4 (1.8)	-0.1
	IEP	3.7 (1.8)	3.3 (1.7)	-0.4
Help [1]	**Total**	**2.4 (1.6)**	**3.5 (2.0)**	**+0.9**
	No IEP	2.3 (1.5)	3.7 (2.0)	+1.4
	IEP	2.8 (2.1)	3.3 (2.1)	+0.5

[12]Self-report variables were measured on a Likert-type scale: [1](1-6); [2](1-5).

Appendix C

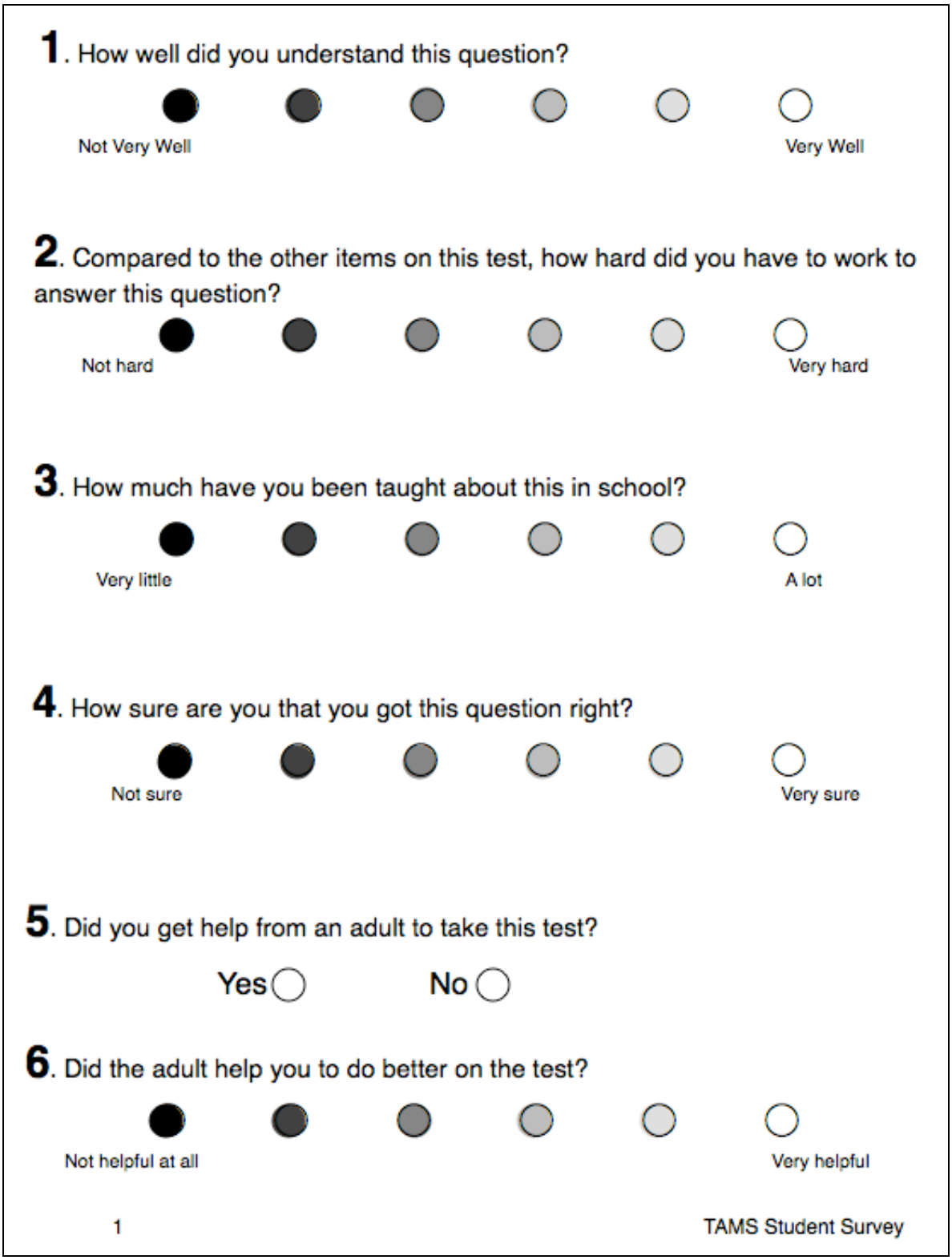


Figure C1. Student post-test survey, sample page.

REFERENCES

American Educational Research Association, American Psychological Association, & the National Council for Measurement in Education. (1999). *Standards for Educational and Psychological Testing* (1999). Washington, DC: Author.

Beattie, S., Grise, P., & Algozzine, B. (1983). Effects of test modifications on the minimum competency performance of learning disabled students. *Learning Disability Quarterly, 6*, 75-77.

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009a). *TAMI Accessibility Rating Matrix.* Nashville, TN: Vanderbilt University. Available at http://peabody.vanderbilt.edu/tami.xml

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2009b). *TAMI Accessibility Rating Matrix Technical Supplement.* Nashville, TN: Vanderbilt University. Available at http://peabody.vanderbilt.edu/tami.xml

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2010). Test accessibility: Item reviews and lessons learned from four state assessments. Submitted to *Educational Measurement: Issues and Practice* for publication.

Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2007). Item Accessibility and Modification Guide. Unpublished rating scale.

Beddow, P. A., Kettler, R. J., & Elliott, S. N. (2008). *Test Accessibility and Modification Inventory*. Nashville, TN: Vanderbilt University. Available at http://peabody.vanderbilt.edu/tami.xml

Bennett, R. E. (2001). How the internet will help large-scale assessment reinvent itself. *Education Policy Archives, 9*.

Bloom, H.S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008, October). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *MDRC* working paper.

Borg, W. R., (1979). Teacher coverage of academic content and pupil achievement. *Journal of Educational Psychology, 71,*, 635-645.

Britton, B. K., Van Dusen, L., Gulgoz, S., & Glynn, S. M. (1989). Instructional texts rewritten by five expert teams: Revision and retention improvements. *Journal of Educational Psychology, 81,* 226-239.

CAST. (2008). Universal design for learning guidelines version 1.0. Retrieved November 26, 2008 from http://www.cast.org/publications/UDLguidelines/version1.html.

Center for Universal Design. (1997). The principles of universal design. Retrieved August 4, 2008 from http://www.design.ncsu.edu/cud

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293-332.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*, 151-170.

Clark, R., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: Evidence-based guidelines to manage cognitive load.* San Francisco, CA: Pfeiffer.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum.

Compton, E., & Elliott, S. N. (2006-2009). *Consortium for Alternate Assessment Validity and Experimental Studies.* USDE Enhanced Assessment Grant (S368A060012). Vanderbilt University.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis, 2*, 7-25.

Dewey, J. (1913). *Interest and effort in education*. Cambridge, MA: Houghton Mifflin.

Ebel, R. L. (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement, 14*, 352-364.

Educational Testing Service. (2009). ETS Guidelines for Fairness Review of Assessments. Retrieved on November 13, 2009 from http://www.ets.org/Media/About_ETS/pdf/overview.pdf

Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M. C., Roach, A. T., & Bolt, D. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children, 76,* 475-494.

Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (2009). Research and strategies for adapting formative assessments for students with special needs. In the *Handbook of Formative Assessment* (H. L. Andrade & G. J. Cizek, Eds.). New York, NY: Routledge.

Feldman, E., Kim, J. S., & Elliott, S. N. (In press). The effects of accommodations on adolescents' self-efficacy and test performance. *The Journal of Special Education*.

Fuchs, L. S., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice, 16,* 174-181.

Garner, R., Alexander, P. A., Gillingham, M. G., Kulikowich, J. M., & Brown, R. (1991). Interest and learning from text. *American Educational Research Journal, 28*, 643-659.

Garner, R., Gillingham, M. G., & White, C. S. (1989). Effects of "seductive details" on macroprocessing and microprocessing in adults and children. *Cognition and Instruction, 6*, 41-57.

Goetz, E. T., & Sadoski, M. (1995). The perils of seduction: Distracting details or incomprehensible abstractions? *Reading Research Quarterly, 30*, 500-511.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*, 181-191.

Graves, M. F., Prenn, M. C., Earle, J., Thompson, M., Johnson, V., & Slater, W. H. (1991). Improving instructional text: Some lessons learned. *Reading Research Quarterly, 26*, 110-122.

Graves, M. F., Slater, W. H., Roen, D., Redd-Boyd, T., Duin, A. H., Furniss, D. W., et al. (1988). Some characteristics of memorable expository writing: Effects of revisions by writers with different backgrounds. *Research in the Teaching of English, 22*, 242-265.

Grise, P., Beattie, S., & Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test modifications make a difference. *Journal of Educational Research, 76*, 35-40.

Haertel, E. H., & Calfee, R. C. (1983). School achievement: Thinking about what to test. *Journal of Educational Measurement, 20*, 119-132.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items.* Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15(3)*, 309-344.

Harp, S. F. & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90*, 414-434.

Herman, J. L., Klein, D. C. D., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice, 4*, 16-24.

Hollenbeck, K. (2002). Determining when test alterations are valid accommodations or modifications for large-scale assessment. In G. Tindal & T. Haladyna (Eds.), *Large scale assessment programs for all students* (pp. 109-148). Mahwah, NJ: LEA.

Hollenbeck, K., Rozek-Tedesco, M., & Finzel, A. (2000, April). *Defining valid accommodations as a function of setting, task, and response.* Presentation at the annual meeting of the Council of Exceptional Children, Vancouver, BC.

Johnstone, C., Liu, K., Altman, J., & Thurlow, M. (2007). Student Think Aloud Reflections on Comprehensible and Readable Assessment Items: Perspectives on What Does and Does Not Make an Item Readable. Technical Report 48. National Center on Educational Outcomes.

Johnstone, C., Thurlow, M., Moore, M., & Altman, J. (2006). *Using systematic item selection methods to improve universal design of assessments* (Policy Directions 18). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Jones, N. L., & Apling, R. N. (2005, May). CRS Report for Congress: The Individuals with Disabilities Education Act (IDEA): Overview of P.L. 108-446. Retrieved from http://assets.opencrs.com/rpts/RS22138_20050505.pdf

Kettler, R. J., Rodriguez, M. C., Bolt, D. M., Elliott, S. N., Beddow, P. A., & Kurz, A. (In press). Modified multiple-choice items for alternate assessments: Reliability, difficulty, and the interaction paradigm. *Applied Measurement in Education*.

Mayer, R. E. & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*, 43-52.

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Education and Policy Analysis, 17*, 305-322.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81-97.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38,* 1-4.

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

Roach, A. T., Beddow, P. A., Kurz, A., Kettler, R. J., & Elliott, S. N. (In press). Using student responses and perceptions to inform item development for an alternate assessment based on modified achievement standards. *Exceptional Children*.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*, 3-13.

Rose, D. H. & Meyer, A. (2006). *A practical reader in universal design for learning*. Boston, MA: Harvard University Press.

Russell, M. (In press). Computerized tests sensitive to individual needs. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of Accessible Achievement Tests* (S. N. Elliott, R. J. Kettler, P. A. Beddow & A. Kurz, Eds.). (1st ed.). Springer.

Schraw, G. (1998). Processing and recall differences among seductive details. *Journal of Educational Psychology, 90*, 3-12.

Thompson, S. J., Johnstone, C. J. & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). Accessibility principles for reading assessments. Minneapolis, MN: National Accessible Reading Assessment Projects.

U.S. Department of Education. (2001). *Elementary and secondary education (No child left behind) act*. Retrieved November 25, 2008 from http://www.ed.gov/policy/elsec/leg/esea02/index.html

U.S. Department of Education. (revised July, 2007). Standards and assessments peer review guidance. Washington, D.C.: Author.

Wade, S. E., Schraw, G., Buxton, W. M., & Hayes, M. T. (1993). Seduction of the strategic reader: Effects of interest on strategies and recall. *Reading Research Quarterly, 28*, 3-24.

Web Accessibility Initiative (2008a). Introduction to understanding WCAG 2.0. Retrieved November 26, 2008 from http://www.w3.org/TR/2008/WD-UNDERSTANDING-WCAG20-20081103/intro.html#introduction-fourprincs-head

Web Accessibility Initiative. (2008b). Web Content Accessibility Guidelines (WCAG) Version 2.0. Retrieved November 26, 2008 from http://www.w3.org/TR/WCAG20.

Winter, P. C., Kopriva, R. J., Chen, C., & Emick, J. E. (2007). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences, 16*, 267-276.